Attention to *Non*-Adopters

Kaitlyn Zhou[1,2,4] **Kristina Gligorić**[1,3] **Myra Cheng**[1] **Michelle S. Lam**[1] **Vyoma Raman**[1,4]
Boluwatife Aminu[1] **Caeley Woo**[1] **Michael Brockman**[1] **Hannah Cha**[1] **Dan Jurafsky**[1]
[1]Stanford University, [2]Together AI,
[3]Johns Hopkins University, [4]Cornell University
kaitlynz@cornell.edu

Abstract

Although language model–based chat systems are increasingly used in daily life, most Americans remain non-adopters of chat-based LLMs — as of June 2025, 66% had never used ChatGPT (Sidoti and McClain, 2025). At the same time, LLM development and evaluation rely mainly on data from adopters (e.g., logs, preference data), focusing on the needs and tasks for a limited demographic group of adopters in terms of geographic location, education, and gender. In this position paper, we argue that incorporating *non-adopter* perspectives is essential for developing broadly useful and capable LLMs. We contend that relying on methods that focus primarily on adopters will risk missing a range of tasks and needs prioritized by non-adopters, entrenching inequalities in who benefits from LLMs, and creating oversights in model development and evaluation. To illustrate this claim, we conduct case studies with non-adopters and show: how non-adopter needs diverge from those of current users, how non-adopter needs point us towards novel reasoning tasks, and how to systematically integrate non-adopter needs via human-centered methods.

1 Introduction

Large language models (LLMs) have made tremendous progress toward supporting humans in everyday tasks. While there is a substantial population of adopters – people who use these systems – there remains a large population of *non-adopters*. For example, as of June 2025, 66% of Americans have never used ChatGPT, the most popular commercial LLM chat model (Sidoti and McClain, 2025).[1] Furthermore, adoption is not random but stratified, with lower adoption rates in historically technologically underrepresented demographic groups (McClain, 2024; Suárez and García-Mariñoso, 2025, i.e., rural, less educated, women).

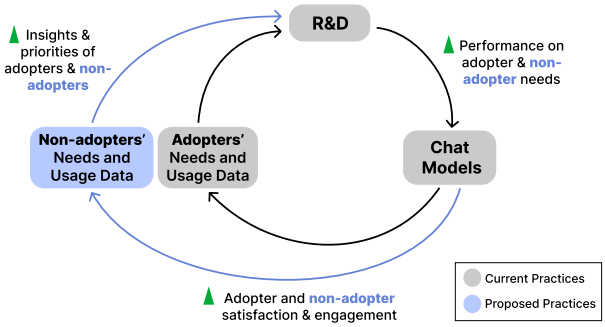

Figure 1: Proposed development cycle with the inclusion of non-adopter needs and preferences.

At the same time, LLM developers lean heavily on interaction logs, preference data, and user studies to study and improve LLMs, including informing us of the current (*and future*) use cases and failures of LLMs (Ding et al., 2022; Pei and Jurgens, 2023; Wang et al., 2022; Zhao et al., 2024; Tamkin et al., 2024; Xiao et al., 2024). Although these methods provide us with convenient insights, they necessarily center on current adopters, overlooking the needs of a broader (*potential*) population.

In this position paper, we argue that **developing broadly useful chat models requires incorporating non-adopters' needs into chat model design, evaluation, and deployment**. We take the position that current LLM development centers on the needs of current users, which represents a narrow demographic group and risks 1) failing to meet the needs of a broad and distinct user audience and 2) failing to advance LLM capabilities on a more diverse set of contexts and tasks.

To illustrate our position, we conduct a set of case studies based on qualitative interviews

[1]**Chat models**: We refer to conversational LLMs with standalone interfaces as chat models (e.g., ChatGPT, Claude, or Gemini), limiting the scope of our research to a technology where users are *intentionally* interacting with it. This definition excludes LLMs that are embedded into other systems like auto-text completion or search summarization.

Findings of the Association for Computational Linguistics: ACL 2026, pages 1336–1366
July 2-7, 2026 ©2026 Association for Computational Linguistics

($n = 23$) and an online survey ($n = 230$) with non-adopters to shed light on the new attitudes, challenges, and unmet needs of non-adopters. Our studies reveal two main findings: (1) rather than being driven by resistance to AI, many non-adopters articulate a tension between their desire to use chat models and the specific difficulties they face in learning how to use new technologies, and (2) tasks prioritized by non-adopters are distinct from those of adopters and are under-represented in chat model evaluations (e.g., navigating basic digital services like healthcare portals, coordinating caregiving responsibilities, accessing contextualized information etc.).

The bias of focusing on existing users has long been studied in HCI through the lens of non-use (Wyatt et al., 2002; Wyatt, 2005; Satchell and Dourish, 2009). In NLP, current adopter-centered practices risk widening the divide between adopters and non-adopters as training datasets, benchmarks, and evaluation metrics are molded to meet the needs of current users. By making this position statement, we hope to direct attention toward the needs of non-adopters and to translate foundational principles from participatory design and inclusive technology adoption to develop a concrete research agenda for the NLP community. We present the following:

1. **Position (§ 2)** We introduce the argument and background for systemically engaging non-adopters.

2. **Why Should We Care? (§3)** We present two case studies ($n = 23, n = 230$) to illustrate why it's important for researchers to consider non-adopters, highlighting that non-adopters are not all active resisters, and that they represent a demographically distinct user audience.

3. **Non-Adopters Point Us Towards Novel Tasks (§4)** We present novel tasks that emerge from non-adopter needs (e.g., navigating basic digital services, managing domestic labor, planning physical tasks), and engage with the NLP literature to illustrate how they differ from current tasks and evaluations.

4. **Practical guidelines (§5)** We identify existing practices that may inadvertently exclude non-adopter perspectives (e.g., 5% of U.S. online crowdworkers have never used chat models, a stark contrast to the 66% observed in the general U.S. population §5.1) and provide recommendations for uplifting non-adopter needs (e.g., re-balancing data annotation and interaction logs, participatory design for developing evaluations, and non-adopter-centered task ideation).

2 Position: Attention to Non-Adopters

Our position is motivated by two key concerns in the adoption and development of chat models: the inequity that arises when language technologies are adopted by only a portion of the population, and the narrow scope of model evaluation when it centers solely on the preferences and tasks of current adopters.

Concern #1: Inequities In Model Access The current development of chat models is at risk of creating a new *digital divide* where some have access to reap the benefits of chat models while others are shut out (Van Dijk and Hacker, 2003). Despite the growing financial cost of using language models (OpenAI, 2025; Anthropic, 2025; Google, 2025), *"access,"* here, is primarily not a physical constraint, but rather a constraint based on the usability and applicability of the technologies. We argue that the current de facto focus on early adopters in data collection and evaluation is not neutral, but rather actively embeds the preferences of a tech-savvy demographic into the foundation of models, which can have far-reaching consequences that may be unknown to us today. A historical parallel comes from automotive safety, where seatbelts were primarily tested on male bodies, and women today continue to experience higher injury and fatality rates as a consequence of this exclusionary design practice (Weiss et al., 2001; Bose et al., 2011). In the context of chat models, we speculate that adopter-centric design risks encoding dominant narratives of how technology "should" be used, perpetuating a cycle where non-adopter needs remain unsupported and their future participation continues to decline (Barocas et al., 2020; Blodgett et al., 2020).

Concern #2: Incomplete Model Evaluations The frequent usage of data from early adopters risks exerting an out-sized and potentially narrow influence on the way chat models are developed. Section 3 provides initial evidence of this already happening. User data through preferences, interaction logs, and user studies are commonly used to understand user needs and influence the future direction and development of chat models (Jiang et al.,

2024; Chen et al., 2024; Liu et al., 2025). However, centering development on existing adopters risks overlooking a diverse and distinct group, and may ultimately lead to an unrepresentative sampling of interactions that doesn't span the full range of real-world human tasks. Furthermore, while chat models developed on current user data *may* display generalist capabilities, if non-adopter needs remain peripheral to chat model development, such opportunistic or incidental uses will lack the systematic design, evaluation, and iteration needed for robust performance.

Proposal We propose that paying attention to non-adopters can address both the burgeoning digital divide and the incompleteness of model evaluations. First, preventing a chat model digital divide is an interdisciplinary challenge requiring both interface-level interventions and foundational changes to model development. Our position, given that we are NLP researchers, will focus on recommendations for NLP practitioners, such as methods, artifacts, and values held in the community. Guided by inclusive design principles (Clarkson et al., 2013), we advocate for the systematic integration of non-adopters' needs at the *outset* of chat model development — shaping datasets, tasks, and evaluations from the beginning. Second, in addressing incomplete model evaluations, the research community has an opportunity to broaden the scope of model evaluation by incorporating non-adopter needs and rigorously testing model performance in new, underrepresented contexts. When future researchers claim *state-of-the-art* capabilities, non-adopter tasks should not be seen as optional, but considered core capabilities that must be robustly evaluated. The integration of non-adopter needs will also trigger additional accountability and evaluation from model providers, holding them accountable to non-adopter use cases (Shen et al., 2021; Lam et al., 2022; Deng et al., 2023).

The History of Studying Non-Use There is a significant literature on the importance and reasons for technological non-use. Researchers have highlighted the distinction between voluntary and involuntary non-use (Wyatt et al., 2002) and why non-use reasons should be considered seriously (Satchell and Dourish, 2009). In arguing for paying attention to chat model non-adopters, we extend the influential work of Wyatt (2005) who argues: *"it is essential to consider the role of non-users in the development of large technical systems... rather than focusing only on the changing relationships between system builders and users."* Many non-adopters will have legitimate reasons to resist adopting or to stop using chat models; however, as research practitioners who have the power to design and shape model capabilities, we urge the community to develop technologies where non-use is a *choice*, rather than an inevitable circumstance.

Ethical Concerns Participatory and co-design methods can accommodate non-adopter needs, but when applied uncritically, they can reproduce or amplify existing power asymmetries (Harrington et al., 2019) and become exploitive (Sloane et al., 2022; Cooper and Zafiroglu, 2024b). Since non-adopters tend to be among historically underrepresented demographic groups (McClain, 2024), these methods must be implemented carefully to avoid widening the digital divide. Critical use, here, requires both inclusion and safeguarding against extractive practices.

3 Why Should We Care?

Why is it important for chat model researchers and developers to consider non-adopters? In this section, we illustrate that non-adopters are *potential users* of chat models, but because their *demographics* and *needs* differ from those of current adopters, they need to be explicitly accounted for.

Case Studies Overview To test our main position, we conducted **Case Study 1**: user interviews with non-adopters ($n = 23$, ages 20-67, across nine different U.S. states) and **Case Study 2**: an online survey via Prolific ($n = 230$, 136 non-adopters and 94 adopters) to illustrate how this population is distinct and can provide new perspectives on chat model development.[2] These studies were IRB approved, and all participants were paid at least $15 USD an hour. For reproducibility, see §A for recruitment details, interview scripts, participant demographics, and survey.

Generalizability of Qualitative Studies Unlike quantitative approaches, our use of qualitative methods is not intended to produce universal claims about the diverse and multifaceted group that is

[2]**Non-adopters**: Defined here as participants who rarely (i.e., once every couple of months) or never use chat models versus **adopters** who use chat models a few times a week or every day. Those who use chatbots *"a few times a month"* are excluded, creating a clear separation between those who have nearly no exposure to chat models from those who have integrated this tool into their everyday lives.

non-adopters. Rather, our aim is to build intuition and map the landscape of who non-adopters are, including how they engage with, resist, or remain ambivalent toward chat models. Our sample size ($n = 23$) is comparable to prior qualitative studies in NLP and HCI (Vaithilingam et al., 2022; Zhou et al., 2022a; Kwon et al., 2024a; Taranukhin et al., 2024), which included 24, 18, 12, and 15 participants, respectively. We focus on English-speaking, U.S.-based participants to situate the study within a specific socio-technical context and enable coherent analysis. While this scope limits generalizability, it offers a foundation for future work across different languages, cultures, and regions.

3.1 Non-Adopters as Potential Users

First, we must recognize that non-adopters make up the **majority** of the U.S. population. Work in June 2025 from the Pew Research Center finds that 66% of Americans have never used ChatGPT and 20% have never even heard of it (Sidoti and McClain, 2025). Not only do non-adopters make up a significant population, but more importantly, our findings suggest that many non-adopters may become future users if given the right opportunities and usability. In our interviews, we find that many non-adopters are not committed to avoiding chat models, but rather, are confronted with obstacles that inhibit easy adoption. For example, non-adopters describe the adoption of technology as a painful learning process that requires labor, time, and support from others.

> *"I'm somewhat bad [at] technology... and no one wanted to help me...* [P4]

> *"I'm in the generation... where there was no technology when I started... I definitely have had to learn a lot along the way... it can be a challenge for me, but I mean I love it, and I hate it at the same time."* [P16]

Despite these learning challenges, they still express a desire to learn and adopt chat models if given the opportunity.

> *"I heard about [ChatGPT] through my younger sister and she said she was going to teach me on how to do it. Well, we've never been able to do it. I wish I knew about it."* [P9]

> *"I mean, that's something I would love to explore... I would, you know, love to learn more about... how to use it"* [P2]

Many non-adopters may not use chat models today, but understanding their needs gives us insights into how to design these language models for a broader population. For example, the verbosity of chat models can be a barrier to participants who speak English as a second language. Participants might also struggle with typing and reading on the screen; enabling text-to-speech interactions in a larger range of models would be critical for this population. These are interface changes as well as model training changes, as preference tuning and generation sampling will likely need to be modified to achieve this accessibility.

3.2 Non-Adopters as Demographically Distinct

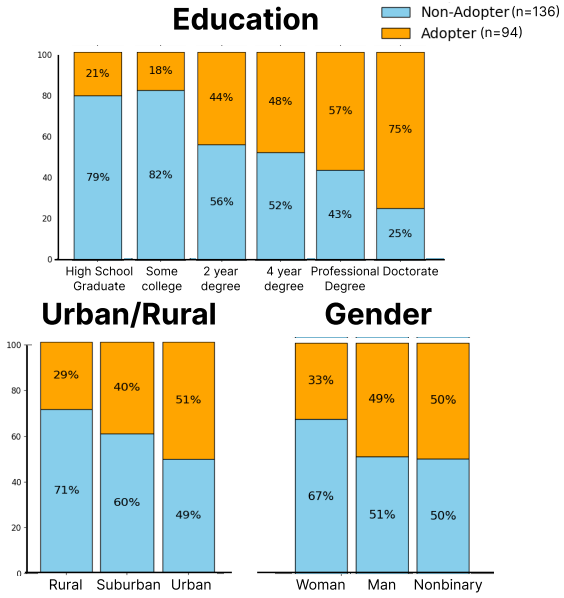

Figure 2: Education, location, and gender demographics of non-adopter and adopters. Non-adopters tend to have less than a 2-year degree, reside in rural or suburban areas, and be women. Adopters tend to have at least a 4-year degree, reside in urban areas, and are more likely to be male. **"Less than high school"* and *"Prefer not to respond"* had 1 response or fewer and were omitted.

Paying attention to non-adopters allows us to design more **equitable** language technologies. The field of artificial intelligence has a long history of introducing systemic, allocational, and representation harms that impact historically underrepresented technology users (Bolukbasi et al., 2016; Schiebinger, 2014; Caliskan et al., 2017; Sheng et al., 2019; Sap et al., 2019; Blodgett and O'Connor, 2017; Zhou et al., 2022b; Santurkar et al., 2023; Blodgett et al., 2016; Jurgens et al., 2017; Koenecke et al., 2020; Ògúnrèmí et al., 2023). Adoption disparities for chat models intersect with model-level biases to reinforce the exclu-

sion of groups already marginalized in technology use and design.

Prior studies have documented demographic gaps in digital technology adoption, and the usage of chat models appears to follow similar patterns (McClain, 2024; Suárez and García-Mariñoso, 2025). Our survey ($n = 230$) confirms these trends: current chat model adopters tend to be highly educated, urban, and predominantly male. We compared adoption rates across demographic groups using a two-proportion z-test, where we test for the rate of adoption. We find that participants without two-year degrees are nearly three times less likely to be adopters compared to those with professional degrees, z = -3.96, $p < 0.001$, or doctorates, z = -3.99, $p < 0.001$; participants who live in rural areas are 40% less likely to be adopters than those living in urban area, z = -2.16, $p < 0.05$, and women are 30% less likely to be adopters compared to men, z = -2.71, $p < 0.01$, Figure 2.[3]

These findings highlight that the disparities between chat model adoption are not arbitrary and can be partially explained by demographic characteristics known to be underrepresented in technology access (Afzal et al., 2023; Elena-Bucea et al., 2021). Here, we stress the urgency to pay attention to current non-adopters who represent a historically underrepresented group of technology users and mitigate the allocation harms resulting from chat models. Hence, although chat models and generative AI more broadly are touted to reduce equity gaps (Fu et al., 2025; Pierson et al., 2025; Gabriel, 2024; Nixon et al., 2024), this is not possible without directly considering and accommodating their non-adopters.

4 Non-Adopters Reveal New Task Priorities

Non-adopters are not only demographically distinct, but they shed light on (1) real-world tasks that are missing in chat model evaluations and (2) ways that current instantiations of classic tasks like question-answering and technology navigation may be misaligned to the needs of a broader population.

4.1 Currently Prioritized Tasks

Our online survey (Figure 3) highlights that adopters highly prioritize writing/reading, technology navigation, and creative tasks — common tasks in the NLP literature (Hadi et al., 2023; Kaddour et al., 2023; Naveed et al., 2024; Shao et al., 2024; Gero et al., 2022; Yao et al., 2022; Fereidouni and Siddique, 2024; Zhu et al., 2023; Chen and Ding, 2023; Tian et al., 2024; Gómez-Rodríguez and Williams, 2023). These findings suggest a potential pattern of self-beneficial technology development known as "me-search" (Bradley and Nash, 2011), which occurs when researchers prioritize the needs of those who are similar to them, in this case, technology developers and current LLM adopters. Even with good intentions, the resulting technologies may end up primarily benefiting those already advantaged, leading to a rich-get-richer effect. Given the known socioeconomic, racial, and gender disparities among computer scientists (Pearl et al., 1990; Lunn et al., 2021), this trend can further exacerbate existing inequalities in LLM benefits.

4.2 Under-prioritized Tasks

In contrast, over half of the non-adopters we interviewed mentioned tasks related to domestic labor. We compared task priorities across adopters and non-adopters using a two-proportion z-test, testing whether participants selected each task as one of their top three most important tasks. Non-adopter survey participants were over four times more likely to choose physical tasks as an important task, $z = 6.13$, $p < 0.001$, and 60% more likely to choose caregiving as an important task compared to adopters, $z = 2.35$, $p < 0.05$. Examples include caring for younger and older children, facilitating family conversations, or supporting a loved one with physical disabilities. Many of these tasks have been recognized in the economics literature to be historically (Sayer, 2005) and currently (Petts et al., 2021) performed by minoritized groups. Our interview data also reveals that these **tasks are not just physical, but also include the invisible cognitive labor known as the mental load** (Dean et al., 2022):

> *"[As] the mom you negotiate and [you] make sure everybody's happy... that's your main focus in life"* [P20].

> *"I always struggle because you have to coordinate with the different people, with the teacher, the decorator, make sure everybody is on the same page... keep it within budget"* [P8].

> *"[Scheduling] is a huge nightmare... If [a chat model] could know all the schedules... plug that in, and then it would tell you... 'This is the best time to see... a possible mixed group of these*

[3] Differences are less pronounced among demographics like race, age, and income, Figure 14.

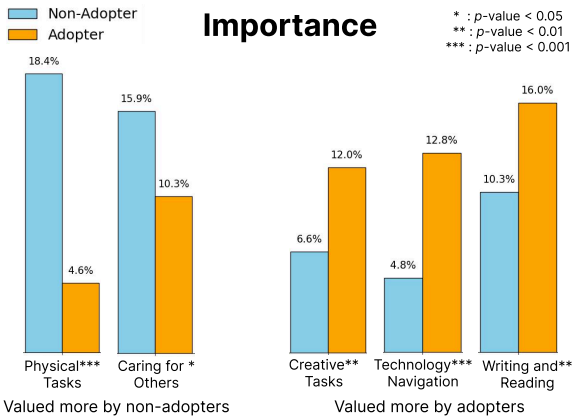

Figure 3: Task importance rankings between adopters and non-adopters with the biggest differences (denoted by % of responders). Non-adopters tended to prioritize physical tasks and caring for others compared to adopters. Adopters tended to prioritize creative tasks, technological navigation, and writing/reading tasks.

> *2 grades of students' — that would be a huge headache reliever."* [P6]

Physical and domestic tasks are rare in the NLP literature and when they appear, they often prioritize technological innovation and are less situated in addressing the real-world needs or requests of specific user groups (Harashima et al., 2016; Sato et al., 2016; Li et al., 2022a; Morales-Garzón et al., 2021; Budzianowski et al., 2018; Liao et al., 2019; Inaba et al., 2022). For example, prior work on developing LLM as a scheduler, a need articulated by participant #6, is in the context of systems optimization rather than scheduling tasks for people and their contexts (Oh et al., 2024). Paying attention to non-adopter needs allows us to see the importance of cognitive labor for physical and domestic tasks and recognize their under-prioritization in language model evaluations.

4.3 Misaligned Tasks

Listening to non-adopters also reveals how existing tasks in chat model evaluation practically differ from what non-adopters need. The interviewees named trying to obtain high-quality information on complex topics as a key pain point. Often, they seek *non-trivial* information that is specific to their unique circumstances. Our case study participants struggled particularly with insurance policies, immigration, and personal finance, and they often need personalized information unique to their identity or circumstance.

> *"As a Black woman...being made aware like what to do in certain situations...diseases that are [prevalent] in like the African American population and how we [can] reduce those type of things from like occurring"* [P5].

> *"I can't even find anyone who really works with people who get pensions...I wish I could get a consultation, someone who understood where we're at"* [P16].

Additionally, non-adopters articulate technology barriers infrequently found in NLP literature, e.g., difficulty typing, submitting documents online, or navigating health care websites. These interaction challenges are ubiquitous and highly limiting, and unaddressed needs could result in users abandoning their tasks entirely.

> "*I've been on [this] BSN program [for the past 5] years because I was not really comfortable with the computer and everything. I quit. I didn't finish"* [P1].

> *"I've been here for almost 2 years and I'm still having trouble finding a dentist... And they have all this online scheduling... their websites are horrible and don't work"* [P12].

While tasks in information retrieval (IR) and question answering (QA) are widely studied by NLP researchers, they typically focus on factoid questions such as questions from reading passages (Rajpurkar et al., 2016; Joshi et al., 2017), standardized tests (Hendrycks et al., 2020), and mimicking historical search queries (Kwiatkowski et al., 2019; Bajaj et al., 2016; He et al., 2017). Recent research on LLMs as agents continues to follow the pattern of retrieving information through a series of steps, querying various dense but non-contextualized materials (Du et al., 2025; Wan et al., 2025). Paying attention to non-adopters highlights the need to evaluate models on non-trivial, contextualized information seeking tasks. One example of this user-centered contextualized work in NLP is from Taranukhin et al. (2024) who grounds their research in a concrete application — supporting Canadian air travel passengers — and consistently centering the needs of a clearly defined user group.

Similarly, numerous tasks in the LLM literature are dedicated to supporting technology interaction for those who are highly technical. Most salient is code generation (Muennighoff et al., 2024; Chen et al., 2021; Li et al., 2022b; Svyatkovskiy et al., 2020; Chen et al., 2022; Lai et al., 2023), but this also includes plotting (Bubeck et al., 2023; Yang et al., 2024) and computer science research more broadly (Desai et al., 2023; Xiao et al., 2023). Meanwhile, research on web agents has focused

on domains like shopping (Yao et al., 2022; Fereidouni and Siddique, 2024), dining (Zhu et al., 2023), and travel (Sun et al., 2022). Interviews with non-adopters highlight their need for support in navigating technology—similar to technologists, though oriented toward a simpler set of tasks.

4.4 The Importance of Non-Adopters

Our user interviews with non-adopters give a glimpse into the complexities that go into underprioritized tasks and reveal opportunities to develop new types of chat model evaluations. Paying attention to non-adopters' tasks opens up new challenges: *Could language models support non-traditional forms of reasoning? Could language models help with the planning of physical and caretaking tasks? What are the safety risks involved?* Our position is that we must give attention to non-adopters, and we urge the NLP community to recognize the importance of asking and prioritizing the needs surfaced by non-adopters.

5 Practical Guidelines: How Could We Incorporate Non-Adopter Needs?

Natural language processing research has been fundamental to the rapid development and advancement of chat models. Here, we outline specific practices in data annotation, benchmark design, and task ideation that may be inadvertently overlooking non-adopters and how alternative practices can be used. Exclusion of non-adopters mirrors psychology's reliance on western, educated, industrialized, rich, and democratic (WEIRD) populations, where narrow participant pools have long skewed findings and limited their generalizability (Henrich et al., 2010). In an effort to counteract these effects, we propose drawing from participatory design (Schuler and Namioka, 1993) and co-design (Steen, 2013; Sanders and Stappers, 2008), in an effort to surface community priorities, redistribute design power, and mitigate harms when working with marginalized communities (Tseng et al., 2025; Sloane et al., 2022; Harrington et al., 2019; Cruz et al., 2023).

5.1 Up-weighting Non-Adopters in Data

Data annotation practices may also be prioritizing the needs of chat model adopters — further embedding adopter preferences into model training reinforces existing usage patterns. Data annotation for natural language processing has long used crowdfunding platforms like Prolific or Amazon Mechanical Turk to recruit humans to generate data, evaluate generations, gather preferences, and eliminate toxicity in language (Munro et al., 2010; Sabou et al., 2012; Zhao and Zhu, 2014). In September of 2025, we conducted a Prolific experiment with 500 U.S. participants (see §A for details). We found that less than 5% of recruited participants have never used chat models, compared to the U.S. average of 66%, showing a huge bias of adopters as annotators (Sidoti and McClain, 2025). Just as uneven annotator demographics can introduce biases into NLP datasets (Ding et al., 2022; Pei and Jurgens, 2023; Wang et al., 2022), emphasizing adopter preferences and data will likely also introduce adopter-centered biases.

Similarly, analysis of interaction logs has been instrumental to our understanding of chat models (Zhao et al., 2024; Tamkin et al., 2024; Xiao et al., 2024) as it reveals the type of interactions, unexpected edge cases, and frequency of chat model usage. However, interaction logs exclusively reflect data from active and frequent users, thus skewing insights toward those with the resources and motivation to engage early and often, and lacking insights on the needs of potential users and how they might leverage LLM capabilities.

One tested way to counteract adopter bias in data annotation and interaction log analysis can be to up-weight non-adopter preferences and labels. Gordon et al. (2022a) propose *jury learning*, where researchers can designate particular groups of annotators whose inputs are weighted more heavily to reflect community values. In data labeling, non-adopter data contributions can be up-weighted to counterbalance the dominance of adopter perspectives. We applied a similar method when we recruited online participants for our own survey data, randomly filtering out chat model adopters and balancing the participant groups evenly. In interaction data, amplifying the voices of low-frequency users and engaging non-adopters through participatory or co-design methods can reveal novel and diverse model interactions.

5.2 Realigning Benchmarks through Non-Adopter Contexts

At the dataset level, non-adopter needs could be integrated by revisiting existing tasks with the goal of including non-adopter contexts. For example, in factoid question-answering, participatory methods could be used to revisit existing artifacts and

ground questions in non-adopter needs. Examples include: understanding new ideas using familiar concepts (e.g., *explaining new lingo*), explaining technical components in simple terms (e.g., *What is 'AirDrop'?*), and contextualizing niche scenarios (e.g., *understanding pensions for retired teachers*). A concrete example stems from the work of Vatsal et al. (2024) who makes advances in the use of chat models to improve the process of prior authorizations for clinicians (i.e., advance clearance for a medical procedure). Re-contextualization here can simply extend this work to help patients understand what is covered in their insurance policies.

Tasks in technology navigation could also be realigned to meet the needs of non-adopters. Here, tasks that are essential activities like filing taxes with free software, scheduling medical appointments, or uploading files to online learning platforms could be prioritized in addition to tasks that serve advanced users.

5.3 Expanding Task Design

Non-adopter needs can be leveraged to expand existing tasks, using methods like need-finding interviews or focus groups. Today, tasks that were once important in the field of natural language processing (e.g., question-answering, information retrieval, summarization, code generation, mathematical computation) still dominate in the evaluation of *state-of-the-art* models. However, non-adopters express interest in a different set of tasks—such as help with the information needs associated with domestic, care-taking, and physical tasks that are rarely reflected in current evaluation benchmarks. In Table 1, we offer an example of one benchmark that operationalizes the information finding needs non-adopters expressed in §3 — showing chat model development beyond factoid QA and towards complex, grounded reasoning that measurably improves human capability augmentation.

Scenario: Unexpected mobility-limiting event (e.g., sudden illness) requiring adaptation to unfamiliar domains.
User Need: Need to research medical information, coordinate logistics, and revise daily routines.
New Task: *Sensemaking* involves information synthesis, context-sensitive reasoning, and emotional attunement.
Model Input: User profile and contextual data, e.g., *"I'm scheduled for surgery in a week, but until then, I need to modify my routines to minimize discomfort and decrease risk of injury. What are some things I should know?"*
Model Output: Agentic and interactive approach. The model might 1) outline a plan of action, 2) synthesize user input, 3) retrieve relevant information, and 4) develop a teaching plan. This may unfold over multiple turns where the model acts like a tutor—gradually building understanding and adapting to user questions.
Evaluation Metrics: Beyond factual accuracy, evaluation could assess user experience: *Did the user feel in control of the interaction? Did they gain a better understanding of the topic? Have they learned strategies to navigate similar problems in the future?*

Table 1: New Potential Benchmark Task: Sensemaking in Unfamiliar, Time-Sensitive Contexts

6 Discussion and Conclusion

Thinking Beyond Current Evaluation Structures Our work emphasizes that non-adopter tasks require thinking beyond the status quo structures of chat model training and evaluation. Non-adopter tasks are currently less structured and less objective (Guerdan et al., 2024). Training chat models on such tasks will also be challenging, as automated metrics are easier to design for structured tasks. The field requires new methods for evaluation and training, as well as greater end-user involvement, to navigate the disagreement inherent to subjective and user-centered tasks (Fleisig et al., 2024; Plank, 2022; Gordon et al., 2022b).

From Chat Model Capabilities to Human Capability Augmentation The integration of non-adopter tasks into chat model development warrants a paradigm shift away from evaluating chat model capabilities in isolation (*"Can the model name the stages of cancer?"*) and towards evaluating *human* capabilities when using chat models (*"Can a user receive support through a new medical diagnosis?"*). A capabilities-augmentation evaluation paradigm extends the work of situated evaluation (Arzberger et al., 2024), as it argues for end-to-end support across user-defined tasks, rather than siloed subtasks defined by external researchers (McClain, 2025). A focus on human capabilities requires a fundamental shift in the tasks we pose and the evaluation metrics by which we measure success as a field.

In this position paper, we argue for the importance of the systematic integration of non-adopter needs into the development of chat models. We illustrate through two case studies how non-adopters differ in demographics and needs from adopters and how their unique perspectives can offer novel insights for expanding LLM evaluations. Lastly,

we identify how current NLP practices can be expanded to incorporate non-adopter needs through the use of human-centered methods.

7 Limitations

Due to our recruitment strategy for study participants, we exclusively interviewed individuals in the United States. Given that varied geopolitics can create different types of needs, our study does not claim to represent needs that are more prevalent in other areas. Furthermore, we conducted interviews in English, which further restricted our recruitment efforts to English speakers in the U.S. Our interview and recruitment strategies relied on internet access and the ability to use a computer or phone, so all participants had that baseline level of technical literacy despite low usage of chat models.

Although physical needs featured as an interesting finding in our study (Section 3), we did not explicitly collect disability-related demographic information, so our evaluation of participants' disability status stems from our conversations during interviews and may be narrow in scope. We also did not interview any participants who self-identified as non-binary, which means that our findings may not reflect the unique needs of that group.

By conducting our empirical research through a single one-hour virtual interview, it is unlikely that we established a bond with participants that would have occurred with in-person or sustained synchronous contact. That means that sensitive or emotionally taxing needs may not have surfaced in our conversations, reducing the scope of our findings.

Finally, we selected OECD's *The Survey of Adult Skills* to contextualize and augment our qualitative coding of interview themes. We acknowledge that this framework is premised on a theory of labor and value that may not be universally agreed upon. Nevertheless, it has proven a useful tool for categorizing the needs we identified and considering gaps in our analysis.

8 Ethical Concerns

Despite our focus on English-speaking U.S.-based participants, we do not advocate for a solely U.S.-centric approach to need identification and task selection. We studied the needs of one sample population that we were able to recruit. Our findings may not generalize to participants from other parts of the world or even others in the U.S., and we would not expect the needs and pain points we identify here to be the most universally salient. Rather, our work attempts to diversify perspectives around which and whose needs should be met in future LLM development, and we seek to further inspire future attempts to do so with different sample populations.

As the use of chat models continues to rise, so do the risks of LLM hallucinations and potential misuse for both adopters and new adopters. In our work, we did not intend to change participants' daily behaviors and habits or advertise products among non-adopters. Given the known biases and failures known to the NLP literature, we took care to debrief the risks of AI after each demonstration.

Acknowledgments

Anonymous First and foremost, we give thanks to all the participants who took part in our interviews and online surveys, whose insights and responses enabled this research.

This work was made possible with the support, feedback, and review from Sofia Kim, Jimin Mun, Jordan Taylor, Mingqian Zheng, Vasudha Varadarajan, Adam Visokay, Federico Bianchi, James Zou, Yongchan Kwon, Shang Zhu, Katelyn Mei, Dexter Crowley, Lindsay Popowski, Nava Haghighi, and Suzanne Lessard.

This work was supported by NSF grant PTA 1258785-1-QABYJ and the Swiss National Science Foundation (Grant P500PT-211127)

References

Arfa Afzal, Saima Khan, Sana Daud, Zahoor Ahmed, and Ayesha Butt. 2023. Addressing the digital divide: Access and use of technology in education. 3:883–895.

Hunter Akridge, Bonnie Fan, Alice Xiaodi Tang, Chinar Mehta, Nikolas Martelaro, and Sarah E Fox. 2024. "the bus is nothing without us": Making visible the labor of bus operators amid the ongoing push towards transit automation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Anthropic. 2025. Pricing — Claude . Accessed: 2025-09-29.

Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. 2024. Nothing comes without its world – practical challenges of aligning llms to situated human values through rlhf. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):61–73.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Shaowen Bardzell. 2010. Feminist hci: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1301–1310.

Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a feminist hci methodology: social science, feminism, and hci. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 675–684.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2020. Fairness and machine learning. *Recommender systems handbook*, 1:453–459.

Andrew A Bayor, Margot Brereton, Laurianne Sitbon, Bernd Ploderer, Filip Bircanin, Benoit Favre, and Stewart Koplick. 2021. Toward a competency-based approach to co-designing technologies with people with intellectual disability. *ACM Transactions on Accessible Computing (TACCESS)*, 14(2):1–33.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022a. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022b. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184.

Su Lin Blodgett. 2021. Sociolinguistically driven approaches for just natural language processing.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*.

Daniel James Bogiatzis-Gibbons. 2024. Beyond individual accountability:(re-) asserting democratic control of ai. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 74–84.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Dipan Bose, Maria Segui-Gomez, ScD, and Jeff R Crandall. 2011. Vulnerability of female drivers involved in motor vehicle crashes: an analysis of us population at risk. *American journal of public health*, 101(12):2368–2373.

DeMethra LaSha Bradley and Robert Nash. 2011. *Mesearch and re-search: A guide for writing scholarly personal narrative manuscripts*. IAP.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Margaret M. Burnett. 2010. Gender hci: what about the software? In *Proceedings of the 28th ACM International Conference on Design of Communication*, SIGDOC '10, page 251, New York, NY, USA. Association for Computing Machinery.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Jiaxun Cao, Hiba Laabadli, Chase H Mathis, Rebecca D Stern, and Pardis Emami-Naeini. 2024. " i deleted it after the overturn of roe v. wade": Understanding women's privacy concerns toward period-tracking apps in the post roe v. wade era. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Kathy Charmaz. 2014. *Constructing grounded theory*. SAGE.

Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. 2022. Codet: Code generation with generated tests. *arXiv preprint arXiv:2207.10397*.

Honghua Chen and Nai Ding. 2023. Probing the "creativity" of large language models: Can models produce divergent semantic association? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12881–12888, Singapore. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Zizhao Chen, Mustafa Omer Gul, Yiwei Chen, Gloria Geng, Anne Wu, and Yoav Artzi. 2024. Retrospective learning from interactions. *arXiv preprint arXiv:2410.13852*.

P John Clarkson, Roger Coleman, Simeon Keates, and Cherie Lebbon. 2013. Inclusive design: Design for the whole population.

Ned Cooper and Alexandra Zafiroglu. 2024a. Constraining participation: Affordances of feedback features in interfaces to large language models. *arXiv preprint arXiv:2408.15066*.

Ned Cooper and Alexandra Zafiroglu. 2024b. From fitting participation to forging relationships: The art of participatory ml. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–9. ACM.

Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1):3–21.

Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Stefany Cruz, Alexander Redding, Connie W Chau, Claire Lu, Julia Persche, Josiah Hester, and Maia Jacobs. 2023. Equityware: Co-designing wearables with and for low income communities in the us. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Liz Dean, Brendan Churchill, and Leah Ruppanner. 2022. The mental load: Building a deeper theoretical understanding of how cognitive and emotional labor over load women and mothers. *Community, work & family*, 25(1):13–29.

Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding practices, challenges, and opportunities for user-engaged algorithm auditing in industry practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Smit Desai, Tanusree Sharma, and Pratyasha Saha. 2023. Using chatgpt in hci research—a trioethnography. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–6.

Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of annotator demographics on sentiment dataset labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–22.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.

Anca Elena-Bucea, Frederico Cruz-Jesus, Tiago Oliveira, and Pedro Simões Coelho. 2021. Assessing the role of age, education, gender and income on the digital divide: Evidence for the european union. *Information Systems Frontiers*, 23(4):1007–1021.

K. J. Kevin Feng, Inyoung Cheong, Quan Ze Chen, and Amy X. Zhang. 2024. Policy prototyping for llms: Pluralistic alignment via interactive and collaborative policymaking. *Preprint*, arXiv:2409.08622.

Moghis Fereidouni and A. B. Siddique. 2024. Search beyond queries: Training smaller language models for web interactions via reinforcement learning. *Preprint*, arXiv:2404.10887.

Casey Fiesler, Shannon Morrison, and Amy S. Bruckman. 2016. An archive of their own: A case study of feminist hci and values in design. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 2574–2585, New York, NY, USA. Association for Computing Machinery.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Batya Friedman. 1996. Value-sensitive design. *interactions*, 3(6):16–23.

Biying Fu, Abdenour Hadid, and Naser Damer. 2025. Generative ai in the context of assistive technologies: Trends, limitations and future directions. *Image and Vision Computing*, 154:105347.

Sonja Gabriel. 2024. Generative ai and educational (in) equity. In *International Conference on AI Research*.

Katy Gero, Alex Calderwood, Charlotte Li, and Lydia Chilton. 2022. A design space for writing support tools using a cognitive process model of writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 11–24, Dublin, Ireland. Association for Computational Linguistics.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Google. 2025. Google AI Plans. Accessed: 2025-09-29.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022a. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, page 1–19. ACM.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022b. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Sophie Grimme, Susanna Marie Spoerl, Susanne Boll, and Marion Koelle. 2024. My data, my choice, my insights: Women's requirements when collecting, interpreting and sharing their personal health data. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Luke Guerdan, Hanna Wallach, Solon Barocas, and Alexandra Chouldechova. 2024. A framework for evaluating llms under task indeterminacy. *Preprint*, arXiv:2411.13760.

Muhammad Usman Hadi, qasem al tashi, Rizwan Qureshi, Abbas Shah, amgad muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. 2023. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects.

Jun Harashima, Michiaki Ariga, Kenta Murata, and Masayuki Ioki. 2016. A large-scale recipe and meal data collection as infrastructure for food research. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2455–2459, Portorož, Slovenia. European Language Resources Association (ELRA).

Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–25.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Long-Jing Hsu, Janice K Bays, Katherine M Tsui, and Selma Sabanovic. 2023. Co-designing social robots with people living with dementia: Fostering identity, connectedness, security, and autonomy. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 2672–2688.

Michimasa Inaba, Yuya Chiba, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. 2022. Collection and analysis of travel agency task dialogues with age-diverse speakers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5759–5767.

Shomik Jain, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson. 2024. Algorithmic pluralism: A structural approach to equal opportunity. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 197–206.

Tingting Jiang, Qian Guo, Yuhan Wei, Qikai Cheng, and Wei Lu. 2024. Investigating the relationships between dialog patterns and user satisfaction in customer service chat systems based on chat log analysis. *Journal of Information Science*, 50(6):1541–1556.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *Preprint*, arXiv:2307.10169.

Lauren Klein and Catherine D'Ignazio. 2024. Data feminism for ai. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 100–112.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Huisung Kwon, Yunjae Josephine Choi, Sunok Lee, and Sangsu Lee. 2024a. Unveiling the inherent needs:

Gpt builder as participatory design tool for exploring needs and expectation of ai with middle-aged users. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Huisung Kwon, Yunjae Josephine Choi, Sunok Lee, and Sangsu Lee. 2024b. Unveiling the inherent needs: Gpt builder as participatory design tool for exploring needs and expectation of ai with middle-aged users. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-tau Yih, Daniel Fried, Sida Wang, and Tao Yu. 2023. Ds-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning*, pages 18319–18345. PMLR.

Michelle S. Lam, Mitchell L. Gordon, Dana"e Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Florian Leiser, Sven Eckhardt, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2023. From chatgpt to factgpt: A participatory design study to mitigate the effects of large language model hallucinations on users. In *Proceedings of Mensch Und Computer 2023*, MuC '23, page 81–90, New York, NY, USA. Association for Computing Machinery.

Shuyang Li, Yufei Li, Jianmo Ni, and Julian McAuley. 2022a. SHARE: a system for hierarchical assistive recipe editing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11077–11090, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022b. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Lizi Liao, Ryuichi Takanobu, Yunshan Ma, Xun Yang, Minlie Huang, and Tat-Seng Chua. 2019. Deep conversational recommender in travel. *arXiv preprint arXiv:1907.00710*.

Q. Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. 2023. Designerly understanding: Information needs for model transparency to support design ideation for ai-powered user experience. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Yuhan Liu, Michael JQ Zhang, and Eunsol Choi. 2025. User feedback in human-llm dialogues: A lens to understand users but noisy as a learning signal. *arXiv preprint arXiv:2507.23158*.

Stephanie Lunn, Leila Zahedi, Monique Ross, and Matthew Ohland. 2021. Exploration of intersectionality and computer science demographics: Understanding the historical context of shifts in participation. *ACM Transactions on Computing Education (TOCE)*, 21(2):1–30.

Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z Gajos. 2024. Evaluating the experience of lgbtq+ people using large language model based chatbots for mental health support. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Colleen McClain. 2024. Americans' use of chatgpt is ticking up, but few trust its election information.

Colleen McClain. 2025. *How the US Public and AI Experts View Artificial Intelligence: The Public and Experts are Far Apart in Their Enthusiasm and Predictions for AI. But They Share Similar Views in Wanting More Personal Control and Worrying Regulation Will Fall Short*. Pew Research Center.

Andrea Morales-Garzón, Juan Gómez-Romero, and Maria J. Martin-Bautista. 2021. Semantic-aware transformation of short texts using word embeddings: An application in the food computing domain. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 148–154, Online. Association for Computational Linguistics.

Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, and Shayne Longpre. 2024. Octopack: Instruction tuning code large language models. *Preprint*, arXiv:2308.07124.

Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Communications of the ACM*, 36(6):24–28.

Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024a. Counterspeakers' perspectives: Unveiling barriers and ai needs in the fight against online hate. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Jimin Mun, Liwei Jiang, Jenny Liang, Inyoung Cheong, Nicole DeCario, Yejin Choi, Tadayoshi Kohno, and Maarten Sap. 2024b. Particip-ai: A democratic surveying framework for anticipating future ai use cases, harms and benefits. *arXiv preprint arXiv:2403.14791*.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky T Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing

and language studies: the new generation of linguistic data. In *NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*, pages 122–130. Association for Computational Linguistics.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models. *Preprint*, arXiv:2307.06435.

Nia Nixon, Yiwen Lin, and Lauren Snow. 2024. Catalyzing equity in stem teams: Harnessing generative ai for inclusion and diversity. *Preprint*, arXiv:2402.00037.

OECD. 2013. *The Survey of Adult Skills*.

Ihudiya Finda Ogbonnaya-Ogburu, Angela DR Smith, Alexandra To, and Kentaro Toyama. 2020. Critical race theory for hci. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–16.

Tolúlọpẹ́ Ògúnrèmí, Wilhelmina Onyothi Nekoto, and Saron Samuel. 2023. Decolonizing nlp for "low-resource languages": Applying abebe birhane's relational ethics. *GRACE: Global Review of AI Community Ethics*, 1(1).

Hyungjun Oh, Kihong Kim, Jaemin Kim, Sungkyun Kim, Junyeol Lee, Du-seong Chang, and Jiwon Seo. 2024. Exegpt: Constraint-aware resource scheduling for llm inference. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, page 369–384, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2025. API Pricing. Accessed: 2025-09-29.

Aviv Ovadya, Luke Thorburn, Kyle Redman, Flynn Devine, Smitha Milli, Manon Revel, Andrew Konya, and Atoosa Kasirzadeh. 2024. Toward democracy levels for ai. *arXiv preprint arXiv:2411.09222*.

Charlie Parker, Sam Scott, and Alistair Geddes. 2019. Snowball sampling. *SAGE research methods foundations*.

Amy Pearl, Martha E Pollack, Eve Riskin, Elizabeth Wolf, Becky Thomas, and Alice Wu. 1990. Becoming a computer scientist. *Communications of the ACM*, 33(11):47–57.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the popquorn dataset. *arXiv preprint arXiv:2306.06826*.

Richard J Petts, Daniel L Carlson, and Joanna R Pepin. 2021. A gendered pandemic: Childcare, homeschooling, and parents' employment during covid-19. *Gender, Work & Organization*, 28:515–534.

Emma Pierson, Divya Shanmugam, Rajiv Movva, Jon Kleinberg, Monica Agrawal, Mark Dredze, Kadija Ferryman, Judy Wawira Gichoya, Dan Jurafsky, Pang Wei Koh, Karen Levy, Sendhil Mullainathan, Ziad Obermeyer, Harini Suresh, and Keyon Vafa. 2025. Using large language models to promote health equity. *Preprint*, arXiv:2312.14804.

Barbara Plank. 2022. The 'problem' of human label variation: On ground truth in data, modeling and evaluation. *Preprint*, arXiv:2211.02570.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, and Arno Scharl. 2012. Crowdsourcing research opportunities: lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, pages 1–8.

Dawn K Sakaguchi-Tang, Jay L Cunningham, Wendy Roldan, Jason Yip, and Julie A Kientz. 2021. Co-design with older adults: examining and reflecting on collaboration with aging communities. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–28.

Elizabeth B.-N. Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *CoDesign*, 4(1):5–18.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Christine Satchell and Paul Dourish. 2009. Beyond the user: use and non-use in hci. In *Proceedings of the 21st annual conference of the Australian computer-human interaction special interest group: Design: Open 24/7*, pages 9–16.

Takayuki Sato, Jun Harashima, and Mamoru Komachi. 2016. Japanese-English machine translation of recipe texts. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 58–67, Osaka, Japan. The COLING 2016 Organizing Committee.

Liana C Sayer. 2005. Gender, time and inequality: Trends in women's and men's paid work, unpaid work and free time. *Social forces*, 84(1):285–303.

Londa Schiebinger. 2014. Scientific research must take gender into account. *Nature*, 507(7490):9–9.

Bianca Giulia Sarah Schor, Emma Kallina, Jatinder Singh, and Alan Blackwell. 2024. Meaningful transparency for clinicians: Operationalising hcxai research with gynaecologists. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1268–1281.

Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC press.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278, Mexico City, Mexico. Association for Computational Linguistics.

Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–29.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Olivia Sidoti and Colleen McClain. 2025. 34% of u.s. adults have used chatgpt, about double the share in 2023. Pew Research Center, Short Reads.

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

Jessie J Smith, Aishwarya Satwani, Robin Burke, and Casey Fiesler. 2024. Recommend me? designing fairness metrics with providers. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2389–2399.

Marc Steen. 2013. Co-design as a process of joint inquiry and imagination. *Design issues*, 29(2):16–28.

Anselm Strauss and Juliet M Corbin. 1997. *Grounded theory in practice*. Sage.

Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, steering, and queering: Treatment of gender in natural language generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Liangtai Sun, Xingyu Chen, Lu Chen, Tianle Dai, Zichen Zhu, and Kai Yu. 2022. Meta-gui: Towards multi-modal conversational agents on mobile gui. *arXiv preprint arXiv:2205.11029*.

Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1609–1621.

David Suárez and Begoña García-Mariñoso. 2025. On the verge of a digital divide in the use of generative ai? *Telecommunications Policy*, 49(7):102997.

Alexey Svyatkovskiy, Shao Kun Deng, Shengyu Fu, and Neel Sundaresan. 2020. Intellicode compose: Code generation using transformer. In *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pages 1433–1443.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, et al. 2024. Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*.

Maksym Taranukhin, Sahithya Ravi, Gabor Lukacs, Evangelos Milios, and Vered Shwartz. 2024. Empowering air travelers: A chatbot for canadian air passenger rights. *arXiv preprint arXiv:2403.12678*.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas Griffiths, and Faeze Brahman. 2024. MacGyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5303–5324, Mexico City, Mexico. Association for Computational Linguistics.

Emily Tseng, Meg Young, Marianne Aubin Le Quéré, Aimee Rinehart, and Harini Suresh. 2025. "ownership, not just happy talk": Co-designing a participatory large language model for journalism. *Preprint*, arXiv:2501.17299.

Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. 2022. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery.

Jan Van Dijk and Kenneth Hacker. 2003. The digital divide as a complex and dynamic phenomenon. *The information society*, 19(4):315–326.

Shubham Vatsal, Ayush Singh, and Shabnam Tafreshi. 2024. Can gpt improve the state of prior authorization via guideline based automated question answering? In *AI for Health Equity and Fairness: Lever-*

aging AI to Address Social Determinants of Health, pages 147–158. Springer.

Haiyuan Wan, Chen Yang, Junchi Yu, Meiqi Tu, Jiaxuan Lu, Di Yu, Jianbao Cao, Ben Gao, Jiaqing Xie, Aoran Wang, Wenlong Zhang, Philip Torr, and Dongzhan Zhou. 2025. Deepresearch arena: The first exam of llms' research abilities via seminar-grounded tasks. *Preprint*, arXiv:2509.01396.

Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose ai dream? in search of the aspiration in data annotation. *Preprint*, arXiv:2203.10748.

Harold B Weiss, Thomas J Songer, and Anthony Fabio. 2001. Fetal deaths related to maternal injury. *Jama*, 286(15):1863–1868.

Jacob O Wobbrock, Krzysztof Z Gajos, Shaun K Kane, and Gregg C Vanderheiden. 2018. Ability-based design. *Communications of the ACM*, 61(6):62–71.

Sally Wyatt. 2005. Non-users also matter: The construction of. *How users matter: The co-construction of users and technology*, page 67.

SM Wyatt, Graham Thomas, and Tiziana Terranova. 2002. They came, they surfed, they went back to the beach: Conceptualizing. *Virtual society*, pages 23–40.

Tao Xiao, Christoph Treude, Hideaki Hata, and Kenichi Matsumoto. 2024. Devgpt: Studying developer-chatgpt conversations. In *Proceedings of the 21st International Conference on Mining Software Repositories*, MSR '24, page 227–230, New York, NY, USA. Association for Computing Machinery.

Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion proceedings of the 28th international conference on intelligent user interfaces*, pages 75–78.

Zhiyu Yang, Zihan Zhou, Shuo Wang, Xin Cong, Xu Han, Yukun Yan, Zhenghao Liu, Zhixing Tan, Pengyuan Liu, Dong Yu, et al. 2024. Matplotagent: Method and evaluation for llm-based agentic scientific data visualization. *arXiv preprint arXiv:2402.11453*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757. Curran Associates, Inc.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Yuchao Zhao. 2024. Design with rural-to-urban migrant women: Opportunities and challenges in designing within a precarious marriage context in south china. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Yuxiang Zhao and Qinghua Zhu. 2014. Evaluation on crowdsourcing research: Current status and future direction. *Information systems frontiers*, 16(3):417–434.

Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. 2024. Beyond preferences in ai alignment. *Philosophical Studies*, pages 1–51.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022a. Deconstructing nlg evaluation: Evaluation practices, assumptions, and their implications. *arXiv preprint arXiv:2205.06828*.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2022b. Richer countries and richer representations. *arXiv preprint arXiv:2205.05093*.

Zichen Zhu, Liangtai Sun, Jingkai Yang, Yifan Peng, Weilin Zou, Ziyuan Li, Wutao Li, Lu Chen, Yingzi Ma, Danyang Zhang, et al. 2023. Cam-gui: A conversational assistant on mobile gui. In *National Conference on Man-Machine Speech Communication*, pages 302–315. Springer.

A Methods Details

Using a bottom-up approach, we begin by outlining the methods that led to our preliminary findings, which form the foundation of our position. User interviews were conducted in the spring of 2024, and an online survey was designed and released in the spring of 2025.

Interview Design Our interview questions were designed to surface participants' needs in their work and personal lives. We draw on *The Survey of Adult Skills*, developed by the Organization for Economic Co-operation and Development (OECD) which surveyed over 166,000 adults (OECD, 2013) to identify six task categories: technology, cognitive, interaction, learning, organization, and physical tasks (OECD, 2013). Participants were asked about their personal and work experiences with these tasks in addition to open-ended questions which were added for completeness, Table 4.

In total, we interviewed 23 participants (Table 5) with limited to no experience with chat models within the United States using snowball sampling via targeted emails (Parker et al., 2019), starting with community members from the authors (e.g.,

Age	#	Gender	#
18-24	22	Woman	115
25-34	47	Man	108
35-44	43	Non-binary	6
45-54	40	No response	1
55-64	60		
65 years or older	18		
Ethnicity	**%**	**Income**	**#**
American Indian/Alaska Native	2%	Under $30k	38
Asian or Asian America	7%	$30-$40k	30
Black or African American	12%	$40 - $50k	13
Hispanic or Latino/a	10%	$50 -$60k	22
Native Hawai'ian/Pacific Islander	<1%	$60-$70k	17
White or European	67%	$70-$80k	20
Other	1%	$80-$90k	10
No Response	<1%	$90-$100k	16
		Over $100k	58
		No response	5
Area of Living	**#**	**Highest Education**	**#**
Rural	41	High school	38
Suburban	128	Some college	38
Urban	61	2 year degree	18
		4 year degree	85
		Professional degree	37
		Doctorate	12
		No response	1

Table 2: Online survey demographics. We had a total of 136 non-adopters (38 who have have never used chat models, 98 who used chat models every few months) and 94 adopters (who used chat models daily). Race demographics are reported in percentages due to multi-racial participants.

in Delaware, Oklahoma, Georgia, California).[4] We then used a bottom-up approach rooted in grounded theory to qualitatively code the interviews (Corbin and Strauss, 1990; Strauss and Corbin, 1997; Charmaz, 2014), yielding 21 unique needs which we refined, categorized, Table 6.

Online Survey Design Informed by our user interviews and literature review, we construct an online survey to understand at a large-scale, how adopters and non-adopters might perceive the importance and painfulness of tasks. We defined non-adopters as participants who never use chat models or use chat models every few months, and adopters are defined as participants who use chat models daily. Our survey included basic demographic questions and questions about task encounter frequency, task importance, and task painfulness, Figure 4.

Our initial list of tasks was again derived from *The Survey of Adult Skills* with modifications, as shown in Table 3. Creative and caregiving asks were added due to their centrality in the literature review and user interviews, respectively. We asked two questions about task importance and painfulness: participants were asked to select the three most important and painful tasks to their work and personal lives (unordered).

We used Prolific's representative sample recruiting feature to recruit survey participants representative of the U.S. census based on gender, age, and ethnicity. We aimed to recruit a balanced sample of adopters and non-adopters, ultimately having 136 non-adopters and 94 adopters (Table 2).

Prolific Experiment Design We constructed an additional online survey to understand the usage of chat models by Prolific annotators. This survey was IRB approved, and all participants were paid at least $15 USD an hour. We used Prolific's standard sampling without additional filtering to get an unbiased understanding of the chat model usage by Prolific annotators. This survey included basic demographic questions and questions about chat model usage. In total, we recruited 500 participants. See 7 for survey questions.

Literature Review The authors then performed a literature review of the Association of Computational Linguistics (ACL) anthology to identify works that could potentially meet the needs expressed by participants. Although there are extensive works in broader research communities on how various LLMs can be *adapted* to meet a broader set of needs, we situate our literature review in the ACL literature, rather than literature in more applied settings, to review LLM development at its conceptualization, rather than its contextualization. We then mapped non-adopters themes to this literature, clustering tasks into major themes based on how well they address non-adopter needs, resulting in broad categorizations of **missing** and **misaligned** tasks.

Pilot Studies: The Difficulties of User Studies Prior to formal interviews, we conducted user studies with OpenAI's ChatGPT interface ($n = 15$) to build intuition of user needs and to iteratively develop our study protocol. A key insight from these pilots is that *it is very difficult for chat model non-adopters to imagine chat model use cases*. Although demonstrations of how chat models can canonically be used may inspire a new user (e.g., code generation and trivia question answering), it would also bias the user towards thinking that those are the *correct* uses. Instead, we ask participants: What are the greatest pain points in your day-to-day life?

[4]Our seed participants are biased towards author communities, however, the team is highly diverse and our participants also reflect this. Author positionalities [redacted]

ChatGPT Demonstrations The interviews included a short demonstration of OpenAI's ChatGPT, where the authors created screen-recordings of ChatGPT attempting these tasks that may be of interest to non-adopters (e.g., brainstorming outing activities with an older parent, writing absence notes to school, producing grocery lists with food restrictions). Interviewers also showcased would ask for the participants' reactions, brainstorm other use cases, and debrief on the safe usage of chat models.

Participant Recruitment and IRB For the online survey, we designed the survey on Qualtrics, which took about 5 minutes to complete. Participants were recruited via Prolific and used Prolific's representative sample recruiting feature, which allowed us to recruit participants which were representative of the simplified U.S. census based on gender, age, and ethnicity. The additional Prolific experiment was also designed on Qualtrics and took 1 minute to complete. Participants were recruited via Prolific using the standard sampling recruiting feature, which allowed us to recruit participants who are representative of the broader population on Prolific.

For the user interviews, we recruited 23 participants with limited to no experience with chat models within the United States using snowball sampling via targeted emails (Parker et al., 2019), starting with community members from the authors (e.g., in Delaware, Oklahoma, Georgia, California).[5]. We used a pre-interview screener to collect background and demographic information and excluded those who have used chat models over 10 times and automatically included all participants who had never used chat models. Each interviewee received a $30 USD gift card for a 45-60 minute video-conference interview conducted via Zoom. In total, we interviewed 23 participants from 9 different states, ranging from 19 to 67 years old (Table 5).[6] Our participant pool size is considered on par with similar qualitative user studies (Smith et al., 2024; Zhou et al., 2022a; Kwon et al., 2024a; Taranukhin et al., 2024; Vaithilingam et al., 2022) which respectively had 13, 18, 12, 15, and 24 participants. This study was IRB approved, and we obtained informed consent from all participants (see Figures 8, 9, 10, 11 for details).

Qualitative Coding of Themes We then used a bottom-up approach rooted in grounded theory to qualitatively code the interviews (Corbin and Strauss, 1990; Strauss and Corbin, 1997; Charmaz, 2014). We iteratively coded and thematically sorted interview excerpts by looking for relations with or among already assigned codes. We first distributed the interview transcripts across all authors for open coding, debriefed in small groups to identify specific needs, and lastly came together to finalize large thematic clusters. Our qualitative coding processes yielded 21 unique needs, which we refined, categorized, and mapped back to our original taxonomy, prioritizing and uplifting needs that were most frequently voiced. Table 6 presents a high-level summary of the concerns voiced by participants.

B Additional Related Work

Our work builds on a body of research aiming to make AI and NLP technologies more inclusive (Burnett, 2010; Bardzell and Bardzell, 2011; Fiesler et al., 2016; Strengers et al., 2020; Blodgett, 2021; Koenecke et al., 2020; Mun et al., 2024b; Zhou et al., 2022a) by adopting participatory (such as Birhane et al., 2022a; Suresh et al., 2024; Cooper and Zafiroglu, 2024a; Zhi-Xuan et al., 2024; Bogiatzis-Gibbons, 2024; Ovadya et al., 2024) or pluralistic approaches (such as Costanza-Chock, 2020; Birhane et al., 2022b; Klein and D'Ignazio, 2024; Jain et al., 2024). Barriers to equity in LLMs include model biases (Bolukbasi et al., 2016; Schiebinger, 2014; Caliskan et al., 2017; Sheng et al., 2019; Sap et al., 2019; Blodgett and O'Connor, 2017; Zhou et al., 2022b; Santurkar et al., 2023) and performance disparities across languages and ways of speaking (Blodgett et al., 2016; Jurgens et al., 2017; Koenecke et al., 2020; Ògúnrèmí et al., 2023), see §B for more.

Closest to our work are recent efforts to ground domain-specific applications of LLMs in users' needs and barriers. Kwon et al. identified similar needs around non-trivial information retrieval (e.g., finance and health) and unpaid labor from those with limited ChatGPT experience. Ma et al. calls for interventions on LLMs at a more foundational level based on evaluations with LGBTQ+ individuals. Mun et al. similarly finds that applica-

[5]Our seed participants are biased towards author communities, however, the team is highly diverse and our participants also reflect this. See author positionalities in Appendix X (redacted for submission).

[6]We contacted 15 men and 40 women, but far fewer men were eligible for the study given their prior interactions with chat models, consistent with known skews in LLM usage.

tions for personal life and society outweigh tasks currently focused on in AI development. HCI research has a long-standing tradition of integrating the distinct needs of various communities into technology design (Muller and Kuhn, 1993; Schuler and Namioka, 1993; Friedman, 1996), addressing aspects such as gender (Bardzell, 2010; Burnett, 2010; Fiesler et al., 2016; Strengers et al., 2020), race (Ogbonnaya-Ogburu et al., 2020) and (Wobbrock et al., 2018; Bayor et al., 2021). More recently, we see a rise in engagement in the design of AI systems with relevant stakeholders through the process of co-design with unique user groups such as designing with older adults (Sakaguchi-Tang et al., 2021), those with dementia (Hsu et al., 2023), bus drivers (Akridge et al., 2024), gynecologists (Schor et al., 2024), and rural migrant women (Zhao, 2024) etc.

In the fairness community, prior work has called for broadening the participation of LLM design (Suresh et al., 2024; Cooper and Zafiroglu, 2024a; Zhi-Xuan et al., 2024), incorporating pluralist views in AI systems (Klein and D'Ignazio, 2024; Jain et al., 2024), and evaluating the unique risks of technology engagement from minoritized users (Grimme et al., 2024; Cao et al., 2024).

Specific to participation in the design of LLMs, HCI works have focused on designing inclusive UX experiences for AI systems (Liao et al., 2023), participatory design with everyday users on LLM hallucination identification (Leiser et al., 2023), collaborating with domain experts on policymaking for LLM alignment (Feng et al., 2024), studying how AI could help in responding to harmful online content (Mun et al., 2024a), and proposing how lay users can surface problematic machine-generated outputs through day-to-day interactions (Shen et al., 2021).

Task	Given Examples	Original OECD Task Name
Technology Navigation	uploading tax documents, scheduling online appointments	Technology - ICT skills
Obtain and Understand Information	understanding information related to health or politics	Cognitive Skills - Problem solving
Calculation Tasks	budgeting, measurement conversions	Cognitive Skills - Numeracy
Writing & Reading	writing emails, reading documents	Cognitive Skills - Reading & Writing
Learning New Skills	recipes, languages, hobbies	Learning - learning
Cooperating, Coordinating, & Negotiating	shared expenses, planning a family trip	Interaction - co-operation & influencing
Communication Tasks	giving instructions, explaining something	Interaction - co-operation & influencing
Physical Tasks	cleaning, cooking, yard work	Physical - Physical requirement
Caring for Others	emotional support, caring for children, sick person, pet	Added - Finding from User Interviews
Creative Tasks	designing, imagining, crafting	Added - Prevalence in NLP Literature

Table 3: List of tasks that survey participants were asked to about in terms of task frequency, importance, and painfulness.

Question	Task Category
Q1: In about 3 to 5 sentences, tell me about yourself.	Intro Question
Q2: What are some tasks that are tedious to complete that you think take longer than they should?	Technology, Cognitive, Learning, Interaction, Organization, Physical
Q3: What are some tasks that are challenging for you to complete?	Technology, Cognitive, Learning, Interaction, Organization, Physical
Q4: What are some topics that you have difficulty obtaining high-quality information for?	Technology, Cognitive, Learning
Q5: What are a few things that you wish were better explained to you?	Technology, Cognitive, Learning
Q8: What are some tasks you encounter that require reading & writing?	Cognitive
Q7: What are some tasks where you work with numbers & need to make calculations?	Cognitive
Q8: What are some creative tasks that you do as a part of your work or free time?	Cognitive
Q9: What are some challenging tasks in your life that involve interacting w/ others?	Interaction, Organization

Table 4: Interview questions and their respective categories

Age	#	Gender	#
18-19	1	Woman	19
20-29	4	Man	4
30-39	3	Non-binary	0
40-49	6	Prefer not to respond	0
50-59	8		
60-69	1		
Ethnicity	**%**	**U.S. Region**	**#**
American Indian/Alaska Native	9%	Middle Atlantic (NJ,PA)	2
Asian or Asian America	26%	East North Central (WI)	1
Black or African American	22%	East South Central (KY)	1
Hispanic or Latino/a	9%	South Atlantic (DE,FL,GA)	9
Middle Eastern/North African	4%	West South Central (OK)	5
White or European	43%	Pacific (CA)	5
Preferred not to Respond	4%		
Occupation	**#**	**Chat Model Usage**	**#**
Architecture and Engineering	2	None	14
Business and Financial	3	1-5 times	4
Education and Library	3	5-10 times	4
Health Care	7	10+ times[7]	1
Sales and Related	2		
Other	6		

Table 5: Participant demographics as self-reported via pre-interview screener. Race demographics are reported in percentages as counts do not add up to 23 due to multi-racial participants.

○ I do not consent to participating in this study

Prolific ID

What is your Prolific ID?

Please note that this response should auto-fill with the correct ID. If it's blank or incorrect, please fill it out.

Technology Perspectives

Which of these artificial intelligence (AI) chatbots have you **used**?

☐ ChatGPT
☐ Co-Pilot
☐ Claude
☐ Llama
☐ Gemini
☐ CharacterAI
☐ PerplexityAI
☐ Other
☐ None of the above

Do you agree or disagree: I am proficient in using AI tools and technology.

○ Strongly disagree
○ Somewhat disagree
○ Neither agree nor disagree
○ Somewhat agree
○ Strongly agree

Demographics

Survey Page 1/5

Age

○ 18-24 years
○ 25-34 years
○ 35-44 years
○ 45-54 years
○ 55-64 years
○ 65 years or older

How often do you use AI chatbots?

○ Daily
○ A few times a week
○ A few times a month
○ Once every few months
○ Never

Do you agree or disagree: I believe AI is improving **my** quality of life.

○ Strongly disagree
○ Somewhat disagree
○ Neither agree nor disagree
○ Somewhat agree
○ Strongly agree

Do you agree or disagree: I believe AI is improving **others**' quality of life

○ Strongly disagree
○ Somewhat disagree
○ Neither agree nor disagree
○ Somewhat agree
○ Strongly agree

Gender identity

○ Female
○ Male
○ Nonbinary/non-conforming
○ Prefer not to respond

What best describes your race or ethnicity?

☐ American Indian or Alaska Native
☐ Asian or Asian American
☐ Black or African American
☐ Hispanic or Latino/a
☐ Middle Eastern or North African
☐ Native Hawai'ian or Pacific Islander
☐ White or European
☐ Prefer not to respond
☐ Other

Which state do you currently live in?

Figure 4: Online survey as presented by Qualtrics

What best characterizes the area you live in?

- ○ Rural
- ○ Suburban
- ○ Urban

How many people live in your household (including yourself)?

Survey Page 2/5

What is your highest level of education?

What is your annual household income?

- ○ Under $30,000
- ○ $30,000-$39,999
- ○ $40,000-$49,999
- ○ $50,000-$59,999
- ○ $60,000-$69,999
- ○ $70,000-$79,999
- ○ $80,000-$89,999
- ○ $90,000-99,999
- ○ Over $100,000
- ○ Prefer not to say

What field do you currently or have previously worked in?

Task Ranking

Survey Page 3/5

The rest of the survey will focus on questions regarding the following ten tasks.

How often do you encounter the following tasks in the past three months?

	Never	Once every few months	A few times a month	A few times a week	Daily
Technology Navigation (e.g., uploading tax documents, scheduling online appointments)	○	○	○	○	○
Obtain and Understand Information (e.g., understanding information related to health or politics)	○	○	○	○	○
Creative Tasks (e.g., designing, brainstorming, crafting)	○	○	○	○	○
Calculation Tasks (e.g., budgeting, measurement conversions)	○	○	○	○	○
Writing Tasks and Reading (e.g., writing emails, reading documents)	○	○	○	○	○
Communication Tasks (e.g., giving instructions, explaining something to someone)	○	○	○	○	○
Caring for Others (e.g., emotional support, caring for children, a sick person, or a pet)	○	○	○	○	○
Learning New Skills (e.g., recipes, languages, hobbies)	○	○	○	○	○
Collaborating with Others (e.g., planning a family trip, sharing expenses)	○	○	○	○	○
Physical Labor (e.g., cleaning, cooking, yard work)	○	○	○	○	○
	Never	Once every few months	A few times a month	A few times a week	Daily

Survey Page 4/5

Pick 3 most important tasks, based on importance in your work or personal life.

- ☐ **Technology Navigation** (e.g., uploading tax documents, scheduling online appointments)
- ☐ **Obtain and Understand Information** (e.g., understanding information related to health or politics)
- ☐ **Creative Tasks** (e.g., designing, brainstorming, crafting)
- ☐ **Calculation Tasks** (e.g., budgeting, measurement conversions)
- ☐ **Writing Tasks and Reading** (e.g., writing emails, reading documents)
- ☐ **Communication Tasks** (e.g., giving instructions, explaining something to someone)
- ☐ **Caring for Others** (e.g., emotional support, caring for children, a sick person, or a pet)
- ☐ **Learning New Skills** (e.g., recipes, languages, hobbies)
- ☐ **Collaborating with Others** (e.g., planning a family trip, sharing expenses)
- ☐ **Physical Labor** (e.g., cleaning, cooking, yard work)

Pick the 3 tasks that have caused **the most pain points** in your work or personal life.

- ☐ **Caring for Others** (e.g., emotional support, caring for children, a sick person, or a pet)
- ☐ **Physical Labor** (e.g., cleaning, cooking, yard work)
- ☐ **Technology Navigation** (e.g., uploading tax documents, scheduling online appointments)
- ☐ **Calculation Tasks** (e.g., budgeting, measurement conversions)

Figure 5: Online survey as presented by Qualtrics

☐ **Obtain and Understand Information** (e.g., understanding information related to health or politics)
☐ **Writing Tasks and Reading** (e.g., writing emails, reading documents)
☐ **Collaborating with Others** (e.g., planning a family trip, sharing expenses)
☐ **Learning New Skills** (e.g., recipes, languages, hobbies)
☐ **Communication Tasks** (e.g., giving instructions, explaining something to someone)
☐ **Creative Tasks** (e.g., designing, brainstorming, crafting)

If you would like to elaborate further on the tasks that have caused the most pain points in your work or personal life, please do so below.

Survey Page 5/5

Given your current resources and responsibilities, how much more assistance do you need for the following tasks?

	None, I can manage this fully on my own	A little assistance would be helpful	Some assistance would be helpful	A lot assistance of would be helpful
Technology Navigation (e.g., uploading tax documents, scheduling online appointments)	○	○	○	○
Obtain and Understand Information (e.g., understanding information related to health or politics)	○	○	○	○
Creative Tasks (e.g., designing, brainstorming, crafting)	○	○	○	○
Calculation Tasks (e.g., budgeting, measurement conversions)	○	○	○	○
Writing Tasks and Reading (e.g., writing emails, reading documents)	○	○	○	○
Communication Tasks (e.g., giving instructions, explaining something to someone)	○	○	○	○
Caring for Others (e.g., emotional support, caring for children, a sick person, or a pet)	○	○	○	○
Learning New Skills (e.g., recipes, languages, hobbies)	○	○	○	○
Collaborating with Others (e.g., planning a family trip, sharing expenses)	○	○	○	○
Physical Labor (e.g., cleaning, cooking, yard work)	○	○	○	○
	None, I can manage this fully on my own	A little assistance would be helpful	Some assistance would be helpful	A lot assistance of would be helpful

Given your current responsibilites, if an **AI tool** was good at the task, how likely would you use it to assist you?

	Extremely unlikely	Unlikely	Neither likely or unlikely	Likely	Extremely Likely	Unsu
Technology Navigation (e.g., uploading tax documents, scheduling online appointments)	○	○	○	○	○	C
Obtain and Understand Information (e.g., understanding information related to health or politics)	○	○	○	○	○	C
Creative Tasks (e.g., designing, brainstorming, crafting)	○	○	○	○	○	C
Calculation Tasks (e.g., budgeting, measurement conversions)	○	○	○	○	○	C
Writing Tasks and Reading (e.g., writing emails, reading documents)	○	○	○	○	○	C
Communication Tasks (e.g., giving instructions, explaining something to someone)	○	○	○	○	○	C
Caring for Others (e.g., emotional support, caring for children, a sick person, or a pet)	○	○	○	○	○	C
Learning New Skills (e.g., recipes, languages, hobbies)	○	○	○	○	○	C
Collaborating with Others (e.g., planning a family trip, sharing expenses)	○	○	○	○	○	C
Physical Labor (e.g., cleaning, cooking, yard work)	○	○	○	○	○	C

	Extremely unlikely	Unlikely	Neither likely or unlikely	Likely	Extremely Likely	Unsu

Do you have any additional comments for us? If you encountered any issues with the survey, please feel free to describe them as well.

Powered by Qualtrics

Figure 6: Online survey as presented by Qualtrics

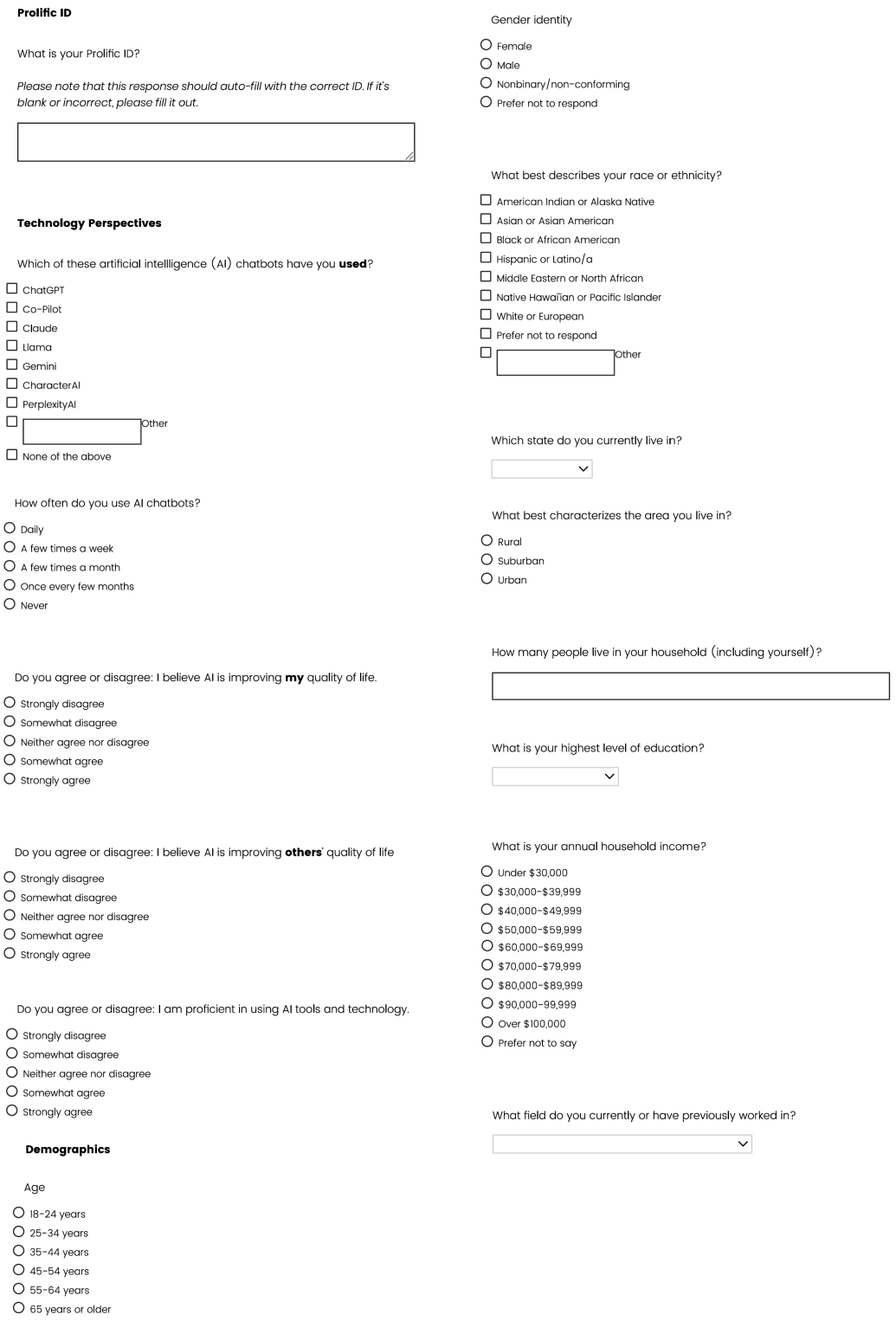

Figure 7: Online Prolific experiment survey as presented by Qualtrics

Recruitment email/post

Subject/Title: Invitation to participate in our research study on Broadening Applications of Public Chat Models

Body: Dear [intended participant],

We are inviting you to be part of our research project about **Broadening Applications of Public Chat Models**. The purpose of the research is to explore applications of chat models for current non-users. You will be asked to 1) complete a pre-survey, 2) engage in a session where you'll be asked questions about your daily activities, 3) shown a tutorial of ChatGPT completing tasks 4) will be asked to complete an optional post-survey. The purpose of this research is to gain insights into how people perceive and use AI. We are specifically engaging with people who have no experience with chat models to learn more about the challenging tasks they encounter and whether chat models can help.

If you participate in our project, you will take part in a short online survey and a short interview (45 - 60 minutes). You will be compensated with a $30 digital Amazon gift card for your time. This research is conducted by the **XXX.**

For more information, please fill out the following form: XXXX
or contact XXXX

Thanks!

Broadening Applications Research Team

Figure 8: Sample Recruitment Email

DESCRIPTION: You are invited to participate in a research study exploring the broader applications of chat models, such as ChatGPT. The study involves understanding people's perspectives, sentiments, and interactions with AI technologies. You will be asked to 1) complete a pre-survey, 2) *engage in a session where you'll be asked questions about your daily activities, 3) shown a tutorial of ChatGPT completing tasks 4) will be asked to complete an optional post-survey. The purpose of this research is to gain insights into how people perceive and use AI.*

TIME INVOLVEMENT: Your participation will take approximately *45 - 60 minutes.*

RISKS AND BENEFITS: The risks associated with this study are *minimal, as the study involves standard interaction with a computer interface and answering survey questions.* Study data will be stored securely, in compliance with XXX standards, minimizing the risk of confidentiality breach. The survey data you submitted prior to this research study will be linked with your interview data but this will all be stored in password safe accounts and accessible only to researchers. The benefits which may reasonably be expected to result from this study include obtaining an opportunity to learn more about chat models and how this tool can be integrated into your daily life. While insights gained from this study may contribute to the broader understanding of AI interaction, we cannot guarantee personal benefits.

PAYMENTS: Upon completion, you will receive $30 USD in the form of an electronic Amazon gift card as payment for your participation in this study.

PARTICIPANT'S RIGHTS: If you have read this form and have decided to participate in this project, please understand your participation is voluntary and you have the right to withdraw your consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. The alternative is not to participate. You have the right to refuse to answer particular questions. The results of this research study may be presented at scientific or professional meetings or published in scientific journals. Your individual privacy will be maintained in all published and written data resulting from the study.

In accordance with scientific norms, the data from this study may be used or shared with other researchers for future research (after removing personally identifying information) without additional consent from you.

CONTACT INFORMATION:
Questions: If you have any questions, concerns or complaints about this research, its procedures, risks and benefits, contact the Protocol Director, **XXX**

Independent Contact: If you are not satisfied with how this study is being conducted, or if you have any concerns, complaints, or general questions about the research or your rights as a participant, please contact the XXXX to speak to someone independent of the research team at XXXX or toll free at XXXX, or email at XXXX You can also write to the XXX.

Indicate Yes or No:

May we contact you about future studies that may be of interest to you?
____ Yes ____ No

Figure 9: Redacted Consent Form

Broadening Applications of Public Chat Models Pre-interview Survey

Thanks for your interest in participating in an interview for our research project about **Broadening Applications of Public Chat Models**.

This information is private and will not be shared with others. Please read through the consent form below and indicate your willingness to participate. If selected for interviews, we will follow up with additional instructions.

* Indicates required question

Study Participation Consent

1. Consent

Please read the attached consent form at the link below and answer the following question.

Mark only one oval.

- ◯ I consent to participating in this study
- ◯ I do not consent to participating in this study

Basic Information

2. Name *

3. Email *

4. Age *

5. Gender identity *

Mark only one oval.

- ◯ Woman
- ◯ Man
- ◯ Non-binary/non-conforming
- ◯ Prefer not to respond
- ◯ Other: _____

6. What state are you from?

7. My race or ethnicity is best described as: *

Check all that apply.

- ☐ American Indian or Alaska Native
- ☐ Asian or Asian America
- ☐ Black or African American
- ☐ Hispanic or Latino/a
- ☐ Middle Eastern or North African
- ☐ Native Hawai'ian or Pacific Islander
- ☐ White or European
- ☐ Prefer not to respond
- ☐ Other: _____

8. Optional: If you'd like to provide additional details on your cultural background, please feel free to do so.

9. What is your level of English proficiency? *

Mark only one oval.

- ◯ 1 – Elementary Proficiency
- ◯ 2 – Limited Working Proficiency
- ◯ 3 – Professional Working Proficiency
- ◯ 4 – Full Professional Proficiency
- ◯ 5 – Native / Bilingual Proficiency

10. What is your occupation? *

Mark only one oval.

- ◯ Architecture and Engineering
- ◯ Arts, Design, Entertainment, Sports, and Media
- ◯ Building and Grounds Cleaning and Maintenance
- ◯ Business and Financial Operations
- ◯ Community and Social Service
- ◯ Computer and Mathematical
- ◯ Construction and Extraction
- ◯ Educational Instruction and Library
- ◯ Farming, Fishing, and Forestry
- ◯ Food Preparation and Serving Related
- ◯ Healthcare Practitioners and Technical
- ◯ Healthcare Support
- ◯ Installation, Maintenance, and Repair
- ◯ Legal
- ◯ Life, Physical, and Social Science
- ◯ Management
- ◯ Military Specific
- ◯ Office and Administrative Support
- ◯ Personal Care and Service
- ◯ Production
- ◯ Protective Service
- ◯ Sales and Related
- ◯ Transportation and Material Moving
- ◯ Prefer not to respond

Figure 10: Pre-interview Screener Survey

11. What is your employment status? *

Mark only one oval.

- Student
- Full-time employee
- Part-time employee
- Prefer not to respond
- Other: _____

Study-Specific Qualifications

12. For research qualification purposes, are you located in the United States? *

Mark only one oval.

- Yes
- No
- Other: _____

13. For compensation purposes, what is your US citizenship status?

Mark only one oval.

- US Citizen
- Permanent Resident
- Prefer not to respond
- Other: _____

14. Have you ever used ChatGPT or other chat-based language models? *

Mark only one oval.

- Yes
- No

15. In the past year, how often have you used ChatGPT or other chat-based language models? *

Mark only one oval.

- None
- 1-5 times
- 5-10 times
- 10+ times

Familiarity with Artificial Intelligence (AI)
For the following questions, please select the option that you feel best describes you.

16. I do not know how AI technology can help me. *

Mark only one oval.

1 2 3 4 5 6 7

Stro ◯ ◯ ◯ ◯ ◯ ◯ ◯ Strongly Agree

17. I can skillfully use AI applications of products to help me with my daily work. *

Mark only one oval.

1 2 3 4 5 6 7

Stro ◯ ◯ ◯ ◯ ◯ ◯ ◯ Strongly Agree

18. It is usually hard for me to learn to use a new AI application or product. *

Mark only one oval.

1 2 3 4 5 6 7

Stro ◯ ◯ ◯ ◯ ◯ ◯ ◯ Strongly Agree

19. I can use AI applications or products to improve my work efficiency. *

Mark only one oval.

1 2 3 4 5 6 7

Stro ◯ ◯ ◯ ◯ ◯ ◯ ◯ Strongly Agree

20. I can evaluate the capabilities and limitations of an AI application or product after using it for a while. *

Mark only one oval.

1 2 3 4 5 6 7

Stro ◯ ◯ ◯ ◯ ◯ ◯ ◯ Strongly Agree

21. I can choose a proper solution from various solutions provided by a smart agent. *

Mark only one oval.

1 2 3 4 5 6 7

Stro ◯ ◯ ◯ ◯ ◯ ◯ ◯ Strongly Agree

22. I can choose the most appropriate AI application or product from a variety for a particular task. *

Mark only one oval.

1 2 3 4 5 6 7

Stro ◯ ◯ ◯ ◯ ◯ ◯ ◯ Strongly Agree

This content is neither created nor endorsed by Google.

Google Forms

Figure 11: Pre-interview Screener Survey (Part 2)

Introduction

[If needed] We are conducting formal interviews as a part of our research agenda and in an effort to remain consistent across our interviews, I'll be partially following a script and may seem more formal than I typically am….

Hi, my name is [x] and thank you so much for joining us today! I'll give a 2-minute overview of the interview and then we'll get started.

I'm an [x] year student and my research interests are focused on artificial intelligence and human-computer interactions. This project is called Broadening Applications and we are trying to research the compatibility between current language technologies everyday user needs. In this interview, we want to learn more about the tasks you have to accomplish in your daily life whether it be related to work or personal life.

There are nine high-level questions with some follow-up questions and then we'll show you a short 2-minute demo and conclude by asking reflective questions. The total interview will take up to 60 minutes. It's helpful for us if you are able to answer in as much detail as possible and there are no wrong answers.. We want to learn about your motivations, your processes, and challenges you face.

As detailed in the consent form, we will be recording this interview to help us with transcribing the interviews. Once our transcriptions are finalized, the video recordings will be deleted. All transcripts will also be anonymized with personal identifiable information removed. You are free to stop this interview at any point during our conversation if you wish to do so. I'll also be taking notes throughout the process and may need a little bit of time in between questions.

Do you have any questions at this point?

Okay great, I'll start the recording and make sure the transcriptions are turned on.

Part 1

1. In about 3 to 5 sentences, tell me about yourself.
    a. Can you tell me more about X? How did you get started in that and what motivates you to continue pursuing this interest?
    b. And what is your background with technologies and interacting with technologies in your everyday work or personal life?

[Make sure you and the interviewee both feel "settled" before moving on to the next question!]

2. What are some tasks that are tedious to complete, that you think take longer than they should?
    a. [EXAMPLE] finding a new health care provider or scheduling an annual exam.
3. What are some tasks that are challenging for you to complete?
    a. What is something that you've been meaning to do for a while?
    b. How do you approach this task right now?
    c. What happened the last time you tried to do it?
4. What are some topics that you have difficulty obtaining high-quality information for?
    a. What are the barriers that prevent you from learning more about it?
    b. Are there tools that could accelerate your learning?
    c. [EXAMPLE] Finance, blockchain, retirement funds
5. What are a few things that you wish were better explained to you?
    a. [EXAMPLE] gerrymandering or intersectionality. It's a fuzzy concept that you wish you had better understanding of
6. [If not already encountered] What are some tasks you encounter and require reading and writing? What are the current difficulties that you encounter?
7. [If not already encountered] What are some tasks where you work with numbers and need to make calculations? What are the current difficulties that you encounter?
8. What are some creative tasks that you do as a part of your work or free time?
9. What are some challenging tasks in your life that involve interacting with others?
    a. Particularly in the realm of cooperating, coordinating, negotiating.
    b. What about tasks around supporting others? This could include teaching, advising, or caring for others?
10. So we've been talking about challenging and tedious tasks [summarize a few of the tasks we discussed]. With these tasks in mind, how likely on a scale of 1 to 5 (being the highest) are to use a tool like chat models like ChatGPT to help accomplish any of the tasks mentioned before?

Part 2: Demo

Great! So that's the conclusion of the first part of our interview. Next, we will show you a short demo of some of the tasks we've seen language models aptly complete. We hope that this demo will help you build a better understanding of the mechanics and features of the tool to help you potentially imagine how it could apply to tasks.

Demo prompts (select the most relevant examples to their tasks)

physical (any), intellectual (any), social (any)	• Help me find a time for the following people to meet based on these schedules: Mark (free 5pm-12am Berlin Time Mon-Thurs), Cathleen (free 12-3pm PT), Benji (free 1-3pm PT and 5-7pm PT)
physical (any), intellectual (any), social (any)	• Write a message to apologetically cancel my Friday night plans with my best friend in a very casual way in a text message.
intellectual (information gathering/learning)	• What are the things I need to know if I want to do a kitchen renovation for under $10k?
intellectual (conceptualization/learning)	• Will you explain to me what a language model is? • Can you put it into simpler terms, like you're explaining it to a middle schooler.
intellectual (reading/writing)	• Will you summarize this document about applying for a visa and put the requirements into bullet point form?
intellectual (numeracy)	• If I have a stipend of [$5500] to spend over nine weeks this summer in San Francisco, excluding housing, come up with a proposed budget for me. I am 32 and want to explore the outdoor community.
intellectual (creativity/content creation	• Give me a list of rainy day activities that I can do with my 55 year old mom and dog, please be really creative and specific.
intellectual (planning and organizing)	• Write a short grocery list for a family of four. • We are vegans. Can you update the list?
social (interaction)	• Here is an email I wrote, revise it: • Dear Principal Watkins, my kid Danielle Parker will be missing school on Monday, May 15 because she is sick. • Make an email subject.
social (support)	• What is a creative way I can teach my toddler about the benefits of sharing?

11. Okay great! So we'd love to hear your reaction to this demo.
    a. What are your initial impressions, thoughts, questions?
    b. What are some tasks where you might use this tool?
    c. What are some tasks where you wouldn't use this tool?
    d. What are some challenges you think you might encounter with this tool?
12. Same question again, how likely (on a scale of 1 - 5) are you to use a tool like chat models like ChatGPT in your everyday life? What are your hesitations about using it in your life?

Debrief:

That's the conclusion of our interview. Thank you so much for your help and thoughtfulness. Here, we're going to debrief very quickly about language models and the potential safety risks. *For example, if I ask a language model for books about [X] (e.g. psychology), it might just output some random book titles that sound real but don't actually exist.* Although our video demonstrates the capabilities of language models, there are numerous cases where the model hallucinates and produces false information. For important information seeking questions, please be vigilant in verifying information.

Figure 12: Interview script read by interviewers. Highlighted parts are asked verbatim.

	TECHNOLOGY		COGNITIVE		INTERACTION		ORGANIZATION		PHYSICAL	
	Software Interaction	Continuous Learning	Non-Trivial Information Retrieval	Non-Needs: Calculations & Creativity	Cross-Cultural Communi-cation	Providing Emotional Support	Domestic Work	Unpaid Labor	Accommo-dations	Physical Chores
P1	X	X		X					X	
P2		X			X					X
P3				X	X					
P4	X	X								
P5			X	X	X					
P6		X		X					X	
P7		X		X		X		X	X	
P8	X	X					X			
P9		X		X						
P10			X	X	X	X	X	X		
P11	X	X		X						
P12			X	X			X			
P13	X				X					X
P14				X		X	X	X		
P15				X			X			X
P16	X		X			X	X	X	X	
P17						X	X	X	X	
P18								X		
P19			X	X						X
P20	X	X		X	X	X	X	X		
P21				X				X		X
P22				X	X			X		
P23				X		X	X	X		
	30%	39%	22%	70%	30%	30%	39%	43%	22%	22%

Table 6: Taxonomy of user needs and occurrences for each participant. **Software interaction**: interacting with existing software systems **Continuous learning**: constantly learning new technologies **Non-Trivial Information Retrieval**: finding or making sense of information unanswerable by search queries alone **Cross Cultural Communication**: digital and real life communication across cultures and generations **Providing Emotional Support**: supporting others **Domestic Work**: domestic duties, including the mental load **Unpaid labor**: historically unpaid labor e.g., organizing, coordinating, planning **Accommodations**: seeking out and learning about physical accommodations **Physical Chores**: routine physical chores **Non-needs:** pain-free cognitive tasks like calculations and creativity

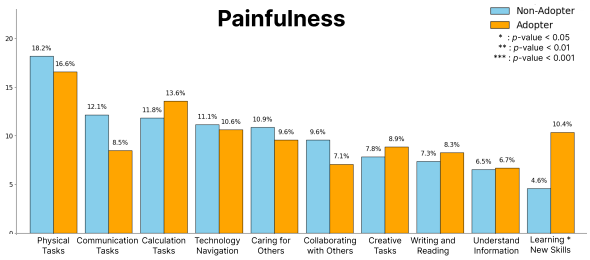

Figure 13: This figure shows task painfulness rankings between LLM adopters and non-adopters. Adopters regarded learning new skills as more painful compared to non-adopters. However, many tasks don't have a significant difference between adopters and non-adopters.

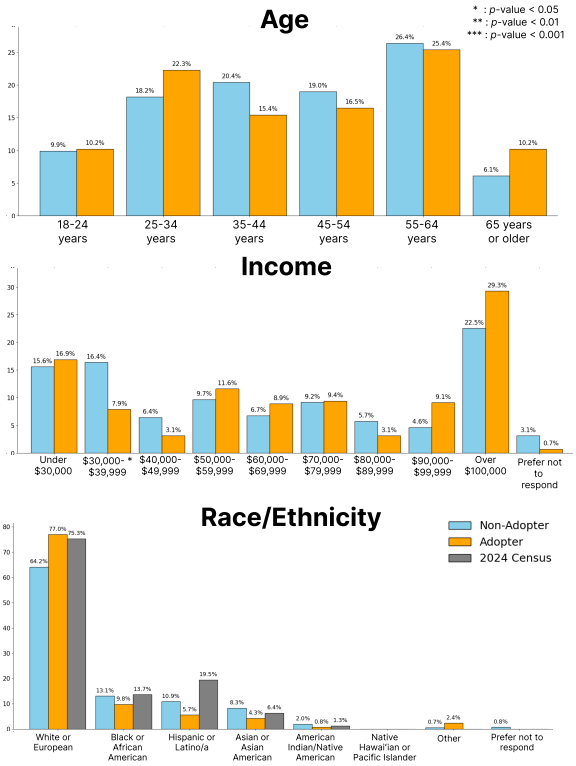

Figure 14: This chart visualizes age, income, and race/ethnicity across non-adopter and adopters. Non-adopters and adopters were similar across age groups. Non-adopters tended to earn under $70,000 annually, whereas adopters tended to be 18-34 years old.

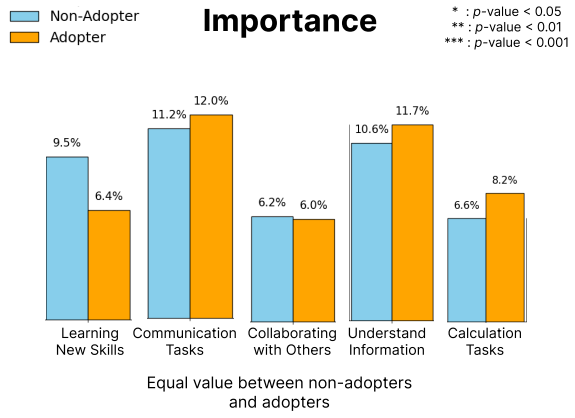

Figure 15: This figure shows task importance rankings between chat model adopters and non-adopters. These tasks show no significant difference in prioritization between adopters and non-adopters.