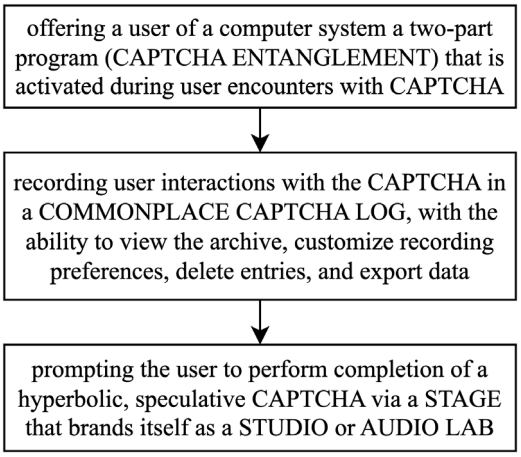
Prepared for Laboratory for Cybernetics (L4C)
May 2025

Institute for Desire Engineering (IFDE)
Represented by Leslie Liu
leslieliu@cmu.edu

CAPTCHA ENTANGLEMENT VIA GENERATIVE AUDIO MODEL



ABSTRACT

Working methods and systems for a two-part application interrogating contemporary CAPTCHA are provided after a broad analysis of questions of locating the ways in which CAPTCHA completers' labor are narrativized and appropriated. The application, positioned here as a critical design proposal, directly engages with gnarly, opaque systems and conceptual provenance in which we find ourselves entangled. Drafts of drawings are enclosed to sketch current working interface considerations; the application comprises a LOG and a STAGE.

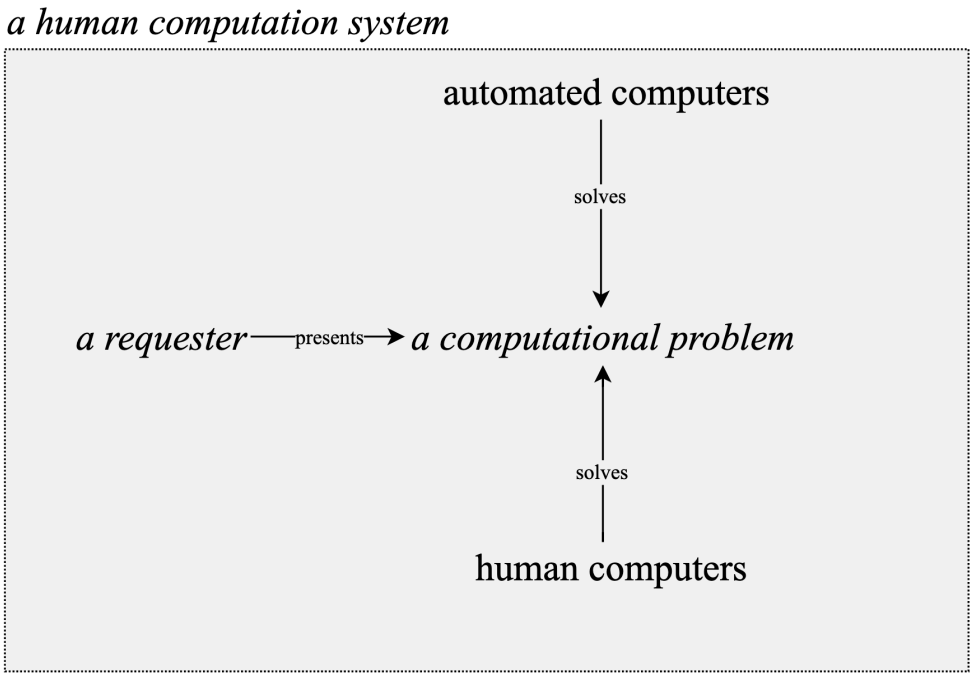

a human computation system
automated computers
solves
a requester
presents
a computational problem
solves
human computers

FIG. 1

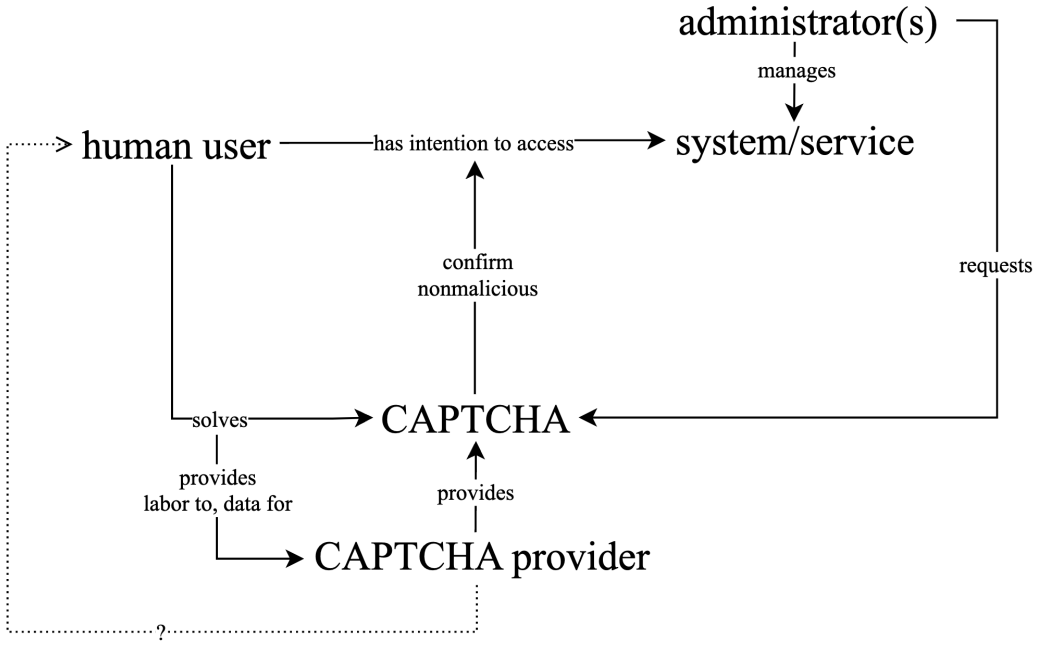

administrator(s)
manages
human user
has intention to access
system/service
confirm
nonmalicious
requests
solves
CAPTCHA
provides
labor to, data for
provides
CAPTCHA provider
?

FIG. 2

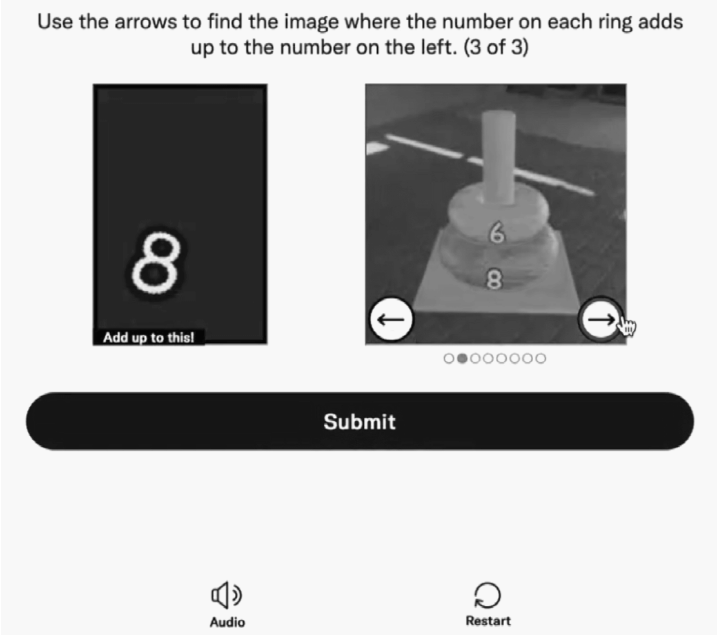


FIG. 3A

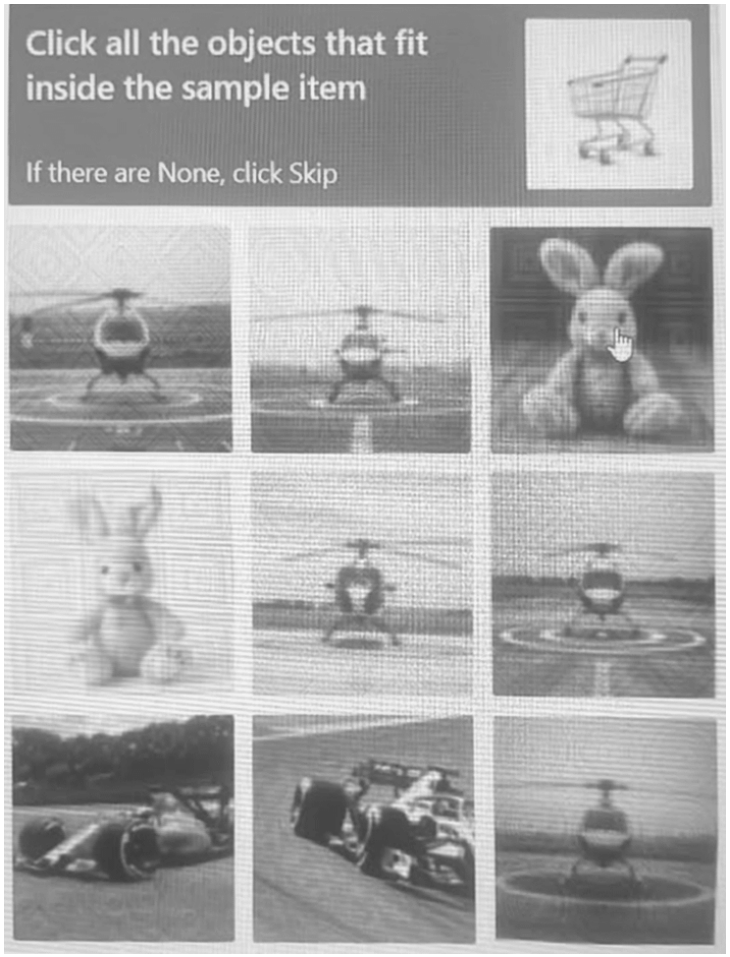
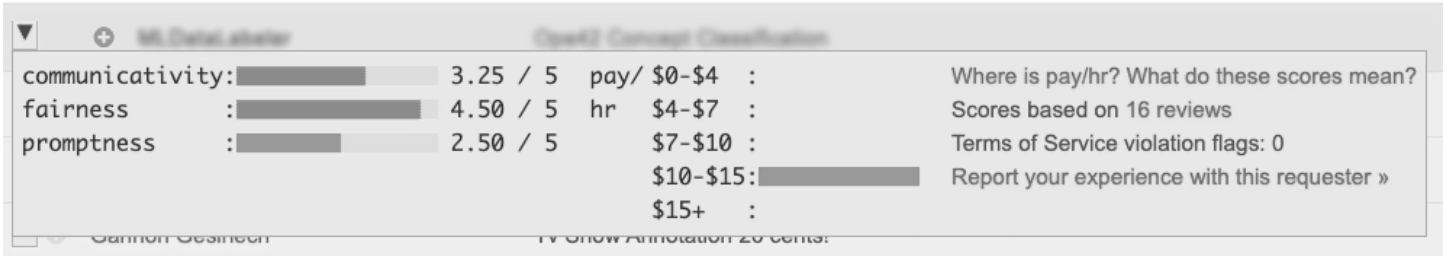


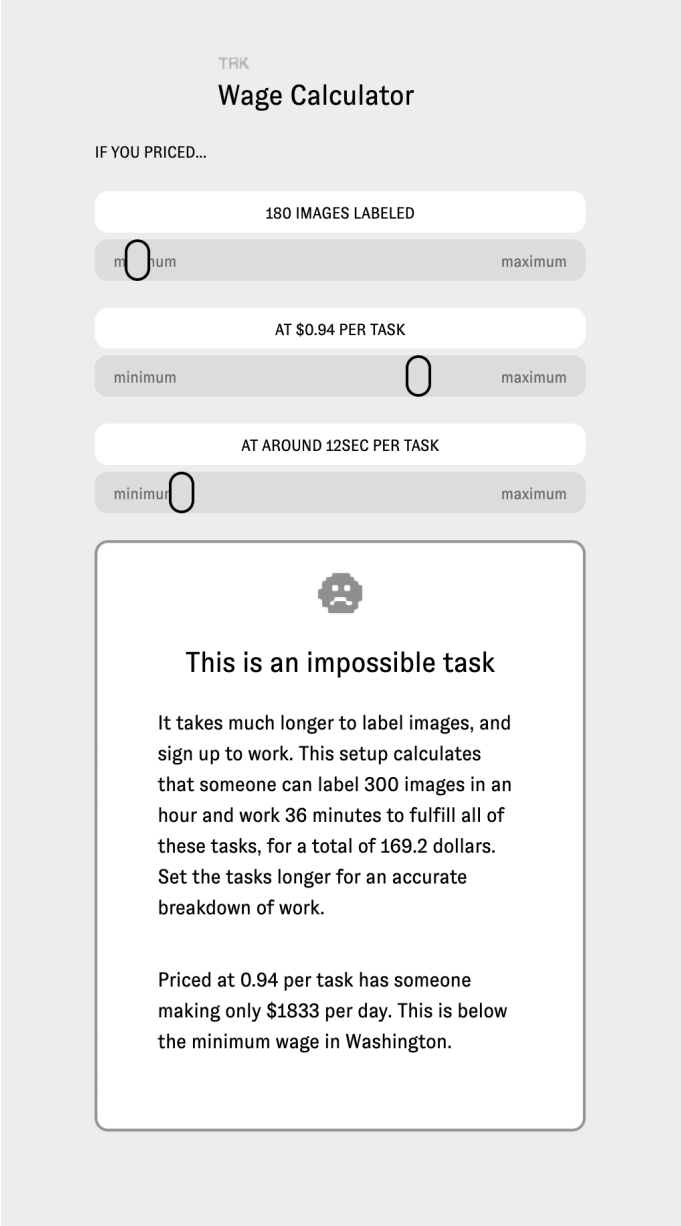
FIG. 3B

via madeleineleplae on Instagram



Irani and Silberman's Turkopticon (partial screenshot above) is a browser plugin/tool for Amazon Mechanical Turk (MTurk) workers (often referred to as Turkers) to hold requesters accountable. The MTurk system is designed such that Turkers overwhelmingly bear the brunt of punitive measures, which often significantly impact their ability to qualify for future work. Turkers use Turkopticon to rate requesters, creating an informal knowledge economy/network for worker power and solidarity.

FIG. 4



A screenshot of Sinders and Diehm's "Technically" Responsible Knowledge: Wage Calculator, a part of Sinders' Feminist Data Set project. Sinders and Diehm cite Turkopticon as an influence on the Calculator, itself an artifact for investigating the supply chains and pipelines underlying machine learning. The Calculator determines whether a data work task is "doable," having consulted creators, Turkers, Fiverr workers, AI artists, and research labs that use MTurk. The Calculator uses U.S. Washington state minimum wage (given Amazon is headquartered in WA) to demonstrate the sheer degree to which data labeling and data training tasks are underpriced.

The authors note the limitation in their methodology, as it approximates "human variants when starting a new task, or the task even changing slightly."

FIG. 5

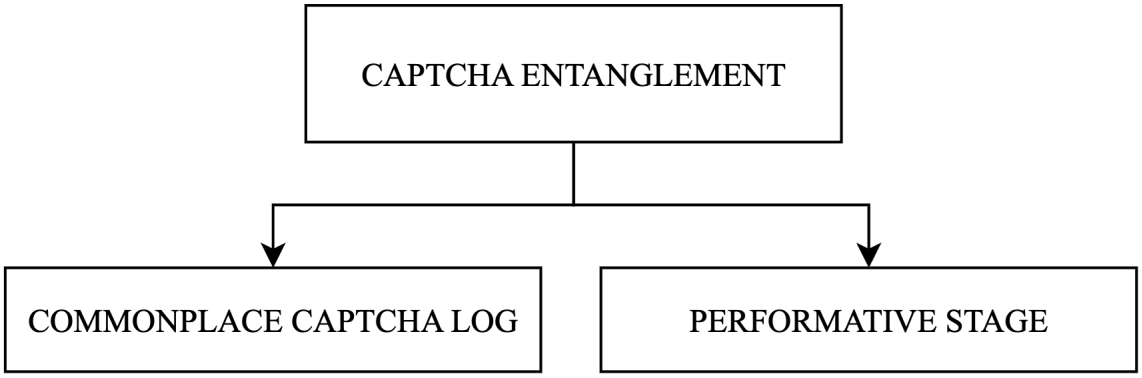

CAPTCHA ENTANGLEMENT
COMMONPLACE CAPTCHA LOG
PERFORMATIVE STAGE

FIG. 6A

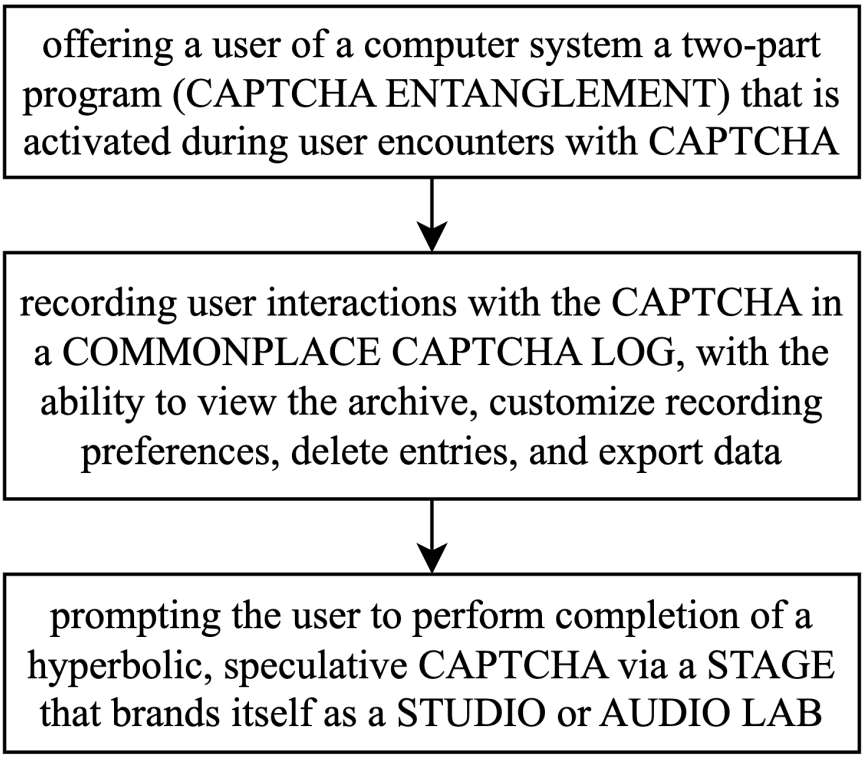

offering a user of a computer system a two-part program (CAPTCHA ENTANGLEMENT) that is activated during user encounters with CAPTCHA
recording user interactions with the CAPTCHA in a COMMONPLACE CAPTCHA LOG, with the ability to view the archive, customize recording preferences, delete entries, and export data
prompting the user to perform completion of a hyperbolic, speculative CAPTCHA via a STAGE that brands itself as a STUDIO or AUDIO LAB

FIG. 6B

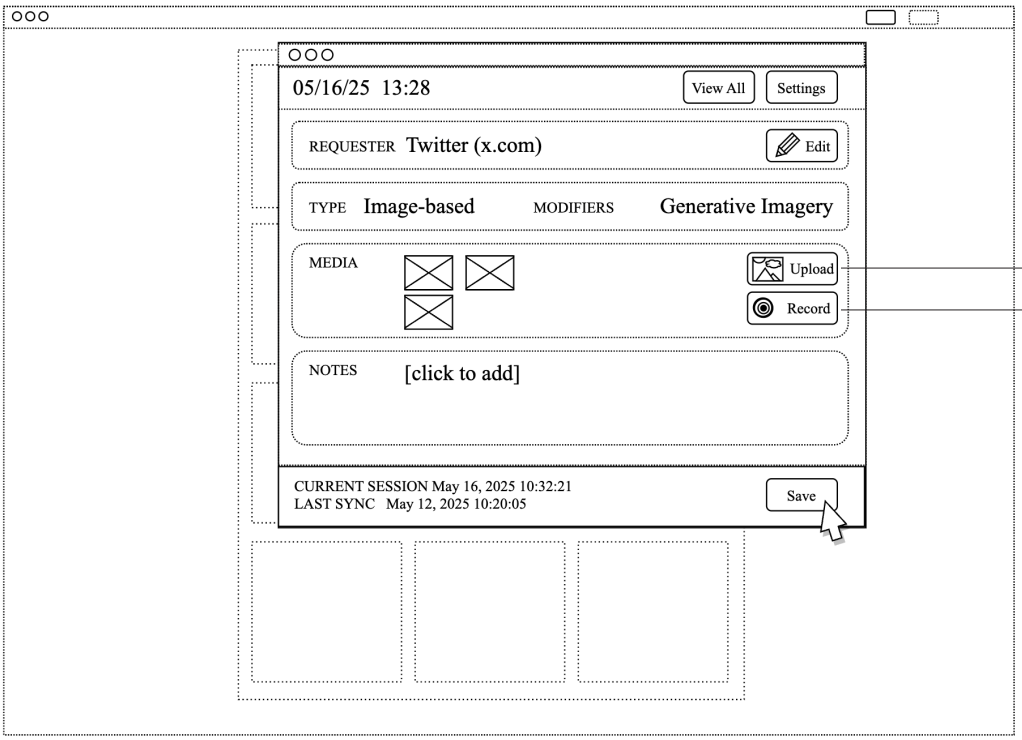

05/16/25 13:28
View All
Settings
REQUESTER Twitter (x.com)
Edit
TYPE Image-based
MODIFIERS Generative Imagery
MEDIA
Upload
Record
NOTES
[click to add]
CURRENT SESSION May 16, 2025 10:32:21
LAST SYNC May 12, 2025 10:20:05
Save
user may upload additional media, or screen-record live

FIG. 7A

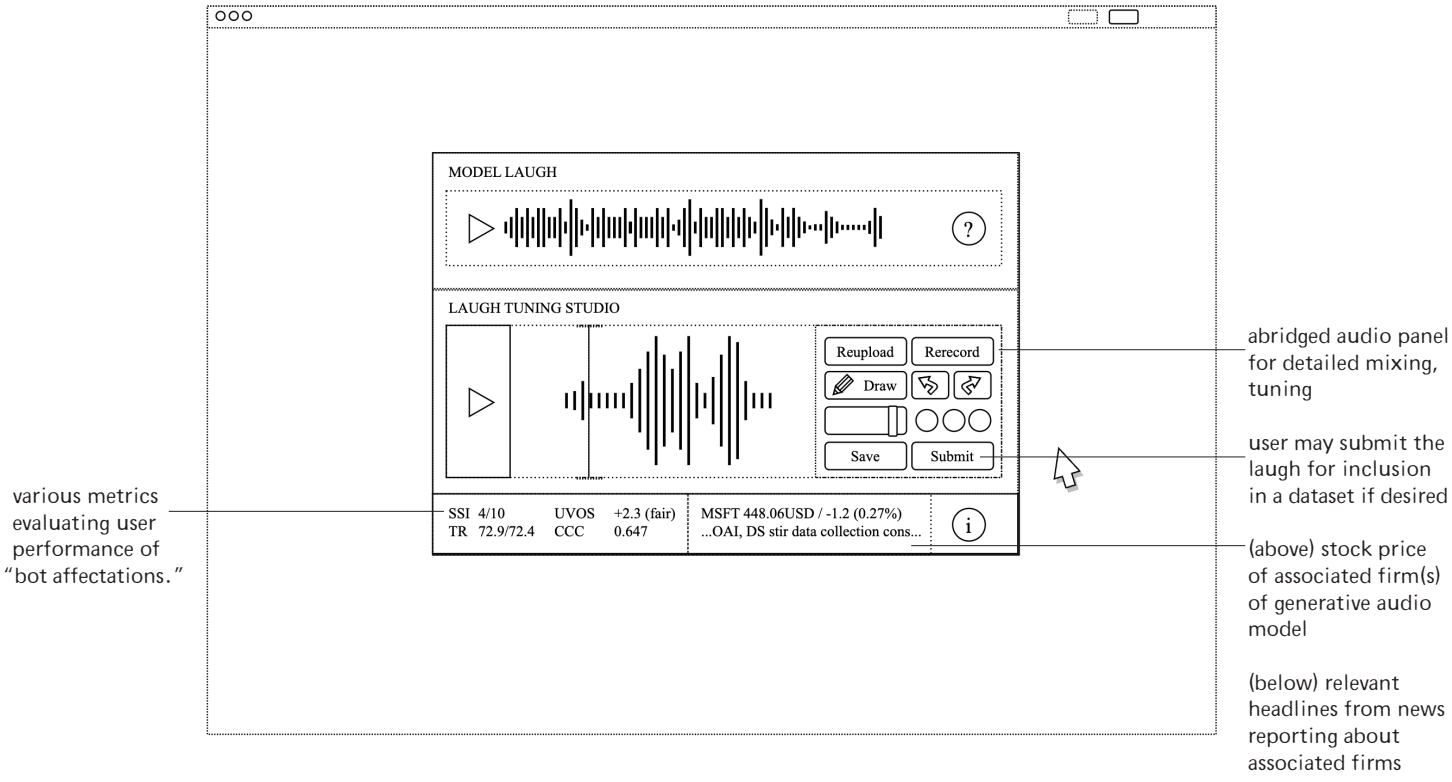

MODEL LAUGH
LAUGH TUNING STUDIO
Reupload
Rerecord
Draw
Save
Submit
SSI 4/10
UVOS +2.3 (fair)
TR 72.9/72.4
CCC 0.647
MSFT 448.06USD / -1.2 (0.27%)
...OAI, DS stir data collection cons...
various metrics evaluating user performance of "bot affectations."
abridged audio panel for detailed mixing, tuning
user may submit the laugh for inclusion in a dataset if desired
(above) stock price of associated firm(s) of generative audio model
(below) relevant headlines from news reporting about associated firms

FIG. 7B

CAPTCHA ENTANGLEMENT VIA GENERATIVE AUDIO MODEL

BACKGROUND

1/ HUMAN COMPUTATION

A CAPTCHA (Completely Automated Turing test to tell Computers and Humans Apart) is an instance of a **challenge-response** test, or **human interactive proof**, used to distinguish discernibly "human" users of a computer system or digital service from their automated, bot[1] counterparts. In cultural retellings of this authentication scheme's historical lineage, CAPTCHA was most notably formalized as a **human computation system** designed to be dual-purpose: by exploiting presumably uniquely "human" computational capabilities, large-scale crowdsourced human labor could thus be harnessed to train and improve machine learning and artificial intelligence (hereafter "AIML") systems, all the while confirming the user as being reliably human. Concurrent academic publishing in the 2000s alternatively situates CAPTCHAs as **Games With A Purpose**.[2]

In their 2011 book Human Computation, Edith Law and Luis von Ahn define human computation as a research area in which "given a computational problem from a requester, [the goal is to] design a solution using both automated computers and human computers." From here follows the definition of human computation (hereafter "hcomp") systems:

> "intelligent systems that organize humans to carry out the process of computation—whether it be performing the basic operations, taking charge of the control process itself, or even synthesizing the program itself." (See FIG. **1** for an illustration.)

Law and von Ahn define a hcomp system's **market** as the workers/human computers involved—"a pool of individuals who are available to work on the computation tasks at hand." While human workers' motivations for participating in this work may vary, a core commonality noted is that "the particular computational problem [they] decide to devote time and effort to help solve has significant value to them[, which] is often more than just monetary." The authors position requesters (entities/bodies/organizations that, in this case, dispatch a CAPTCHA to confirm the user's identity) as

> "also the stakeholders of any human computation [system], whose goal is to solve the computational problem of interests in the most accurate, efficient and economical way possible. **A human computation system is not sustainable without satisfying the needs and wants of both workers and requesters**."

Implicit in the construction of a CAPTCHA is what Luis von Ahn, Manuel Blum, and Benjamin D. Maurer in their 2008 patent term both the "verify" and "read" parts of the scheme, where the former is a controlled experiment whose success allows the system to gain confidence in the user's input into the unknown, "read" part used for training. While this distinction may be more easily spotted in traditional text-based CAPTCHAs, in contemporary implementations of image-based, behavioral, and invisible CAPTCHAs,[3] it is difficult and near impossible **to materially locate in what way(s) and when users are contributing to the hcomp system**. Thus a question arises from how the authors define the "interests'" computational problem—that it must be framed in such a way that its solution is "as accurate, efficient, and economical as possible"—how do we define the expectations of the proposed solutions of using CAPTCHA when not all parties that constitute the "interests" (namely, the human computers) are able to stake a claim in the solution's development?

Granted, this protestation may be swiftly dismissed with the remark that those who participate in the development of the field—academics, industry leaders, among others—are often by default those who have the most say; however even Law and von Ahn note that the layered nature of CAPTCHAs is not always clearly communicated by the CAPTCHA providers, nor is it often concretely understood by the human workers:

> "What most people don't realize is that by typing a CAPTCHA they are also participating in the largest distributed human collaboration project in history."

2/ USER MOTIVATION

While early text-based CAPTCHAs (patented as reCAPTCHA) that leveraged human perception to digitize old print media sported the motto, "Stop Spam. Read Books," current popularly used CAPTCHA schemes do not address the ways in which human computation is used beyond the tacit assumption, as advertised, as necessary input to somehow inevitable fixtures of public web security infrastructure. Indeed, in their 2025 analysis of the economics underlying the production and collection of human-sourced data, Sebastin Santy et al. argue that intrinsic human motivations have not been sufficiently accounted for, and that majoritarian stances on the design of human-generated data collection schemes such as CAPTCHA prioritize quantity, which is "largely driven by process efficiency, often through task fragmentation and parallelization." However, it is this very fragmentation that, in excessive degrees, can "erode intrinsic motivation, leading to long-term declines in quality." As the authors underscore through an examination of a 2023 study by Andrew Searles et al., "ReCAPTCHA [...] remains widely used today but has blurred the line between voluntary participation and coercion, with users reporting frustration and annoyance."

The coercive element is felt through the recasting of human computers' core need to access a system/service through various filters (see FIG. **2** for a coarse mapping). The system/service administrator is a requester who has a computational problem of assessing whether the user who wants access is "human" and thus presumably has no ill intent to abuse the service. The administrator/requester requests a CAPTCHA, provided by a (often third-party) CAPTCHA provider to respond to the user; the user in turn exchanges their human computation for access to the system/service, but **their implicit interaction with the CAPTCHA provider, who is the intermediary between user and system/service, is technically one-sided: while the user agrees to authenticate themself, the CAPTCHA provider is not accountable to them**. The very real, material, and valuable data that the user provides (often without their own being aware) is effectively obscured, taken for granted.

One might argue that the positioning of the user's humanness is thus a rhetorical means of establishing some sense of "intrinsic motivation" such that the user will want to prove that they are well-intentioned. Considering what it really means to tangibly "measure" a positive case of successful CAPTCHA completion, one might note that "humanness" in this hcomp system has a threefold definition: a human is one who is

1. willing to cede their data to the CAPTCHA provider (whether manually entered or passively volunteered via cookies: browsing history/browser environment, page interactions, screen resolution, cursor movement, etc.);
2. can actually pass the test and will put up with any encountered

"failed tests" by repeatedly completing successive iterations of the CAPTCHA; and

3. potentially subscribes to the notion that the test is in some way important to their self-recognition and personhood.

This positions personhood and humanity as a convenient and helpful ideological and emotional framework, skeuomorph, and metaphor through which to conceptualize the hcomp system involved; "humanness" is defined negatively as "not displaying bot-like patterns" and recursively defined through successful interactions with the system, which are all by necessity opaque for security reasons, presumably. **Are overtures to our so-called personhood and humanness, measured by a black-boxed system, motivation enough? Yet it is this exact vagueness, seeped through this very analysis, that one might also defend as being necessary to the functioning of security systems in which people place their trust**. Obscurity is perhaps the point: that the work one does when completing a CAPTCHA is mysterious (so core to the boundaries of "what is human," nestled deep within the subconscious processing of our biologies) and unimportant (trivial, quotidian, unremarkable because of how "easy," how "natural" it should come to us)—we cede personal responsibility, displacing it to the purview a technical system, so that it can be an arbiter, tie breaker, or decision maker for us.

3/ SOCIAL CONTRACT

As a human computation system, CAPTCHA reflects a core software engineering principle of **modularity**, which David Gray Widder and Dawn Nafus—in their 2023 analysis of locating responsibility and accountability in AI supply chains—describe,

"sets the stage for a refusal to accept a relationship between 'us' developers and 'them' technology users, let alone other affected citizens [...] modularity is an epistemic culture [...] that cultivates a capacity to "bracket off" [... which] makes it an everyday form of the modernist fallacy of the separability of society from technology (Latour, 1993), separating code from harms it enables. It is an example of the social organization of ignorance (Proctor and Schiebinger, 2008), where the focus on one thing (the workings of a single portion of code) yields ignorance of another (the activities of other developers and users)."

Nafus and Widder also highlight the consequences of "scale thinking"—a mental and ideological outlook in which

"notions of scale render **'technical systems as commodities that can be stabilized and cut loose from the sites of their production long enough to be exported en masse to the sites of their use'** (Suchman, 2002: 95). They reinforce the distinction between inside and outside a company and create an important site of cutting a technology loose from its creators."

Applied in our context, the CAPTCHA presents individual humanness as something worth working for as part of the market—the hundreds of millions of people von Ahn refers to, each of whom has contributed to the same species-level project of technological advancement—**however there is no way to recognize the effort expended in the work, for it is so individualized, fragmented, made modular**. There is no sense of connection with others in the crowd, no way to locate your labor and the merits that you reap in reward. The CAPTCHA constructs a market across time and space; a workforce that is pliable and can be called upon with great flexibility and ease, and most critically—is free. Moreover, **the social contract inherent in this system positions human labor not as labor—but rather, enthusiastic existential self-reaffirmation within the narrowly defined bounds (defined by whom?) of the system**. (A vague statement that one's labor "also improves [the CAPTCHA provider's] systems" falls short. For instance, representatives of Google's autonomous vehicle venture Waymo have denied speculations that claim the firm uses image-based (re)CAPTCHA data to train their AV systems—a rejection whose persuasiveness is indeterminate given the firm's history of leveraging free services for crowdsourcing efforts, per GOOG-411, a 2007 speech-recognition-based telephone service.[4])

If we attempt to consider a roughly averaged user experience of CAPTCHA completion as a conversation, there is no reciprocity—there is no conversation: there is no grounding to arrive at an agreement of means and goals. One might even go as far as to claim that many people who find themselves new recruits of the market don't realize that **in a system that <u>would</u> meaningfully offer them agency and real, dignified choices, they would be able to at least understand the role they play in this human computation system**.

SUMMARY OF THE PROPOSAL

The present proposal operates under the hypothesis that **being able to tangibly contextualize one's labor and its potential applications will encourage more ownership and motivation to engage with the system through attempts at understanding and mutual legibility**. It includes two parts: Part I, a LOG that serves as a "data diary" of sorts for the user to keep handy as they browse the web in everyday situations; and Part II, a STAGE where laboring to reach a hyperbolic "gold standard" serves as a provocation for questions and discussions about the concerns elaborated thus far.

A note that the present proposal is not a technical redesign of CAPTCHA, though efforts such as the Cloudflare Turnstile have been undertaken to replace CAPTCHAS, primarily for improved security (such that data is not used for ad retargeting) and user experience.

<u>Human computation</u>

Current CAPTCHA schemes employ human perception, sometimes in addition to contextual data from the user's browser, user interactions with the page, and cookies. In their role as a gatekeeper moderating access to a service, CAPTCHA designs in the past have attempted to make the task more fun for users, for example through the use of simple game mechanics involving image semantics or arithmetic problems. While contemporary schemes, as encountered across the web, have intensified in terms of cognitive labor required (see FIG. **3A**, where the CAPTCHA combines image semantics with an arithmetic problem and FIG. **3B**: the CAPTCHA below also asks for spatial understanding and an implied, general cultural awareness), whether the CAPTCHA provider collects data beyond the scope of the authentication process is unclear. The expectation falls on the user to go to the CAPTCHA provider's Privacy Policy, if it is disclosed, to confirm this. For example, hCaptcha—a reCAPTCHA alternative—states in their FAQ that "hCaptcha analysis in the 'invisible mode' may take place completely in the background. Website or app visitors are not advised that such an analysis is taking place if the user is not shown a challenge." While the usability tradeoff from a security perspective may be desirable here, would presenting the user with a statement of when their own behavior and browsing environment is being analyzed, offer them more agency? To crudely analogize surfing the web to navigating a retail store, we might conceptualize our personal interactions with CAPTCHA as loss prevention measures brick-and-mortar stores put in place: CAPTCHA completion is similar to passing through RF (radio frequency) scanning exit gates; however in certain implementations (invisible CAPTCHAs, specifically) the gates are

imperceptible to users.

Understanding that presenting the user with overwhelming amounts of information may exacerbate adverse reactions, recalling principles of ambient design may be helpful: the LOG suggests an informative panel or interface just once at the start of a browsing session or period of time, and appears (or stays hidden) as specified by user preferences.

User motivation

Current embodiments[5] of CAPTCHA fail to disclose their specific molding of "humanness" and "humanity" and assume that the system's definition of normativity is universally aspirational. They also do not spare any attention to clarifying how the human user fits into a broader market of fellow data workers.

The STAGE problematizes this by making explicit what it means to "improve AIML systems" by offering directions for assessing impact via tech reporting and relevant benchmarks; it also sidesteps the question of personhood by instead optimizing for an artificially defined "bot affectation" whose calculation is self-dislosed to the user should they want to interrogate the "gold standard" that they are striving towards.

Social contract

Central to this project is the question of locating labor; due to the fragmented nature of CAPTCHA data work, users cannot materially understand how they are contributing to (which?) AIML systems. Current CAPTCHA embodiments do not disclose the actual flows and exchanges of labor and access that occur at the site of the interface.

The STAGE, through its formulation of the computational problem (reaching an artificially defined standard) and implementation of specific interaction patterns for user input, directly reflects the "assessment" process. Because the primary modality involved is sound and thus inaccessible to Deaf and Hard of Hearing users, the proposal is under development to be available as a purely visual scheme.

The construction of the "gold standard, model laugh" takes up the question of cognitive labor and acclerates it into novel realms of absurdity. Moving away from distinguishing objects per assumed categorical objectivity, the STAGE uses laughter as a benchmark, making explicit a deliberate and enacted "engineering of emotion" to muddy the otherwise presumably clear-cut distinctions and boundaries between human-bot definitions.

Through the use of state of the art OpenAI generative audio models (gpt-4o-transcribe and gpt-4o-mini-transcribe), the STAGE builds on prior audio model research (LibriSpeech from Johns Hopkins University, 2015; wav2vec 2.0 from Facebook AI, 2020; Whisper from OpenAI, 2022) to exploit user-generated output. The 2015 LibriSpeech corpus is based on LibriVox, a public domain audiobook project entirely run by volunteers. While LibriVox audiobooks are originally intended for literary audiences, they have been adopted wholseale in speech processing and AIML research. (LibriVox's Artificial Intelligence page includes a Q&A item that explicitly states the project's noninvolvement in "[selling] its recordings to everyone," though the recordings' being released into the public domain means that such wide-ranging use for AI trainers "without [LibriVox's/users' permission, for free]," is legal.)

Locating the precise training data used in gpt-4o-transcribe is thus reliant on overt efforts to trace through academic and industry research literature, often complicated by webs of references that are more intelligible to those in the field than the typical layperson/human computer. Moreover what makes the use of this corpus attractive to the present proposal is the way in which OpenAI has advertised gpt-4o-transcribe's ability to be noticeably more "lively." **This liveliness attribute, which end users/consumers can tune (however coarsely) through interactions with the ChatGPT "chatbot" application/interface or the OpenAI API, once again reintroduces a normative conception of humanness**. Here it is recircumscribed technologically by virtue of these "voice agents'" role as undeniably anthropomorphic interface metaphors of human user-generative audio model interaction. The proposal seeks to explicate these tensions in order to help contextualize the user's future interactions with mainstream CAPTCHA.

PRIOR ART

The proposal consults piror art—Lily Irani and M. Six Silberman's Turkopticon (2013; a snapshot in FIG. **4**) and Caroline Sinders and Cade Diehm's Tehnically Responsible Knowledge Wage Calculator (2020; see FIG. **5**).

DESCRIPTION OF THE DRAWINGS

FIG. **6A** presents, in patent semantics, the summative structure of the current invention, which comprises two separate, standalone applications, whose rudimentary roles are outlined in FIG. **6B**. FIG **7A** and FIG. **7B** display the basic interface considerations for the LOG and the STAGE, respectively; see the **Appendix** for detailed interaction flow wireframes.

Part I: Commonplace CAPTCHA Log (LOG)

A1.1–3 demonstrate the interaction flow of the CAPTCHA Log.

The Log serves as an intermediary interface, suggested to the user whenever they encounter a CAPTCHA, prompting them for collection. **A1.1** demonstrates ambient, automated collection and **A1.2** manual CAPTCHA collection.

The Log autopopulates associated metadata of the interaction, with additional buttons and controls to upload contextual information and media as the user desires. A "Notes" field offers a secondary, diary-writing functionality should the user want to add personal information about particular associations, pain points, etc. that the CAPTCHA brings to mind.

A1.3 shows the full log of past CAPTCHAs collected, able to be filtered and/or sorted by various criteria. An export option is available.

The Log functions as a notetaking device and journal, while it fails to address the question of nebulous applications of user labor, through sheer accumulation it seeks to make visible the user's solitary efforts.

Part II: Performative Stage for Laugh Tuning (STAGE)

A2 details the user's interactions with the Stage, involving a hyperbolic performance of tuning one's laugh to match a predefined "gold standard" laugh. Successive attempts to edit the user's original audio clip to resemble the model laugh are judged live by the system, which provides the user more detailed statistics that contextualize current failures to conform to the standard (see annotation on FIG. **7B**). The info button attached to the model laugh, when activated, displays the particular prompts and inputs to OpenAI's generative audio model

for construction of the standard. The model laugh here serves as an approximation for "bot affectations," itself a complex technological artifact borne of an intensive system of data collection (via LibriVox and LibriSpeech as previously described), data cleaning/training, and data labeling. All this labor is not directly addressed by—nor is it the primary focus of—the Stage.

CONTRIBUTIONS

The CAPTCHA ENTANGLEMENT (LOG and STAGE) addresses the lack of material understanding and context afforded the user in user-CAPTCHA interactions. By making explicit the ways in which the standard involved in a test is devised—as well as making more concrete the user's own volunteering of data by prompting for input to be uploaded to the system. While the ENTANGLEMENT does not claim to "solve" any practical, technical conundrum, it serves to resurface intrinsic motivations amongst users and to problematize the various cycles in which human labor is appropriated for ends not initially considered, in the use of LibriVox-turned-LibriSpeech-turned-GPT 4o Transcribe audio model.

Ongoing research and development is under way to bring this system to bear, with user trials forthcoming.

NOTES

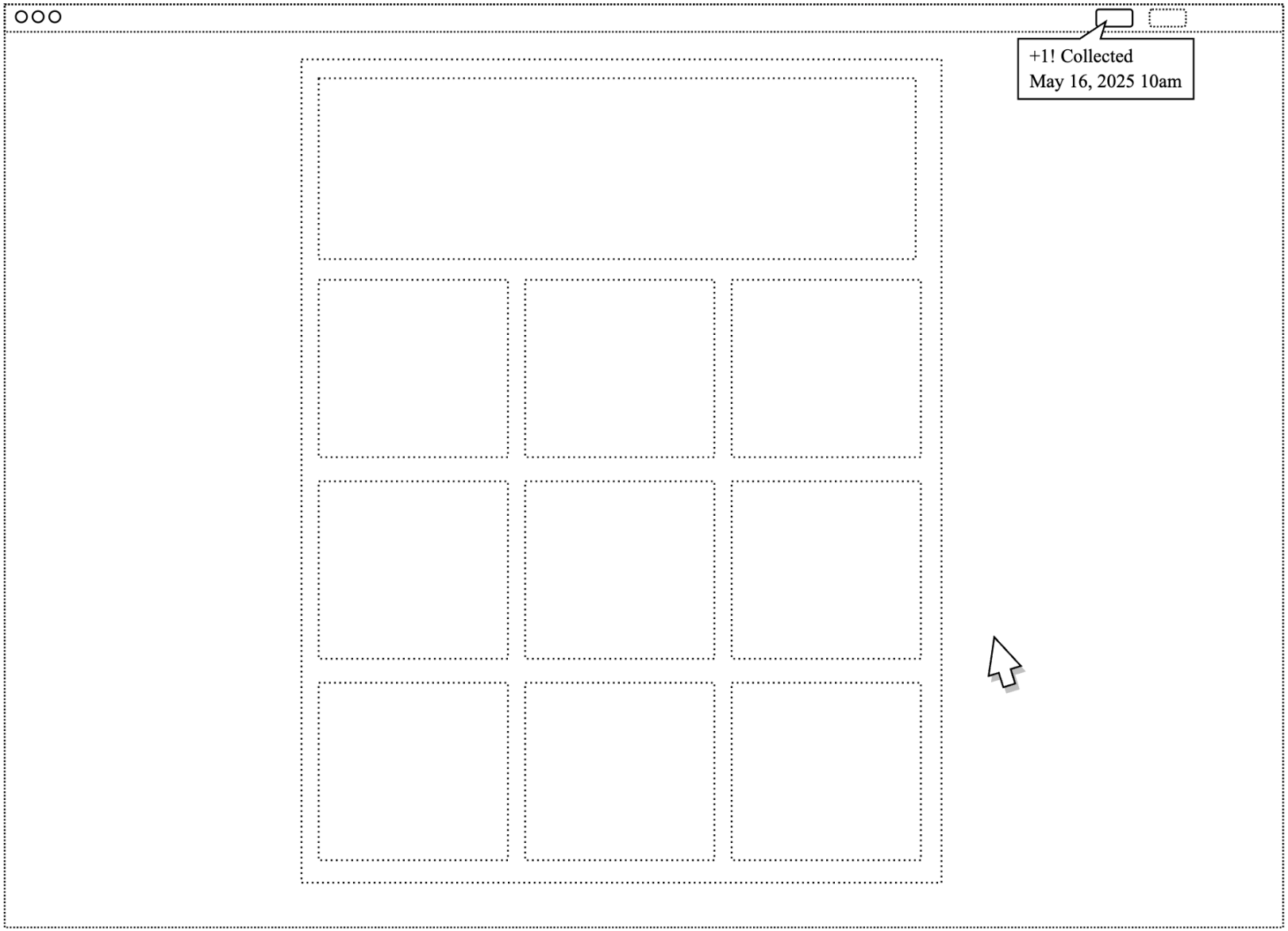
1. The use of "bot" in this report is specific to the following personal definition: "an automated software application." It does not account for the various popular cultural imaginaries surrounding androids and robots.
2. Games With A Purpose was formalized by Luis von Ahn and Laura Dabbish in 2003 and 2008.
3. A 2021 survey of CAPTCHAs by Meriem Guerar et al. titled "Gotta CAPTCHA 'Em All: A Survey of Twenty years of the Human-or-Computer Dilemma" presents ten different groups/categorizations of CAPTCHAs. Behavioral CAPTCHAs may not involve overt user input, relying instead on contextual browser/gestural data. Invisible CAPTCHA refers to the No CAPTCHA reCAPTCHA, aka reCAPTCHA V2, which Google deployed in 2014. No CAPTCHA reCAPTCHA analyzes user behavior information in the background, presenting the authentication process as exceedingly "seamless" and hassle-free. The invisible CAPTCHA recalls ongoing and long-standing discussion in interaction design about the debate over whether "no user interface is the best interface." The current invention detailed in this report takes an argmentative stance in this debate, emphasizing a seamful, overcommunicative attitude.
4. Reporting from 404 Media—"Pokémon Go Data 'Adding Amplitude to War Is Obviously an Issue,' Niantic Exec Says" by Emanuel Maiberg, November 25, 2024—further underscores the lack of accountability or recourse available should labor be appropriated for other ends beyond the user's imagination or consideration.
5. "Embodiment" is a word specific to patents that this document aims to subtly problematize. The patenting impulse, it appears, to externalize, physicalize, and making concrete the abstract (through diagrams and especially through language) is a curious inversion of the dynamic at play in the prevailing CAPTCHA schemes critiqued.

APPENDIX

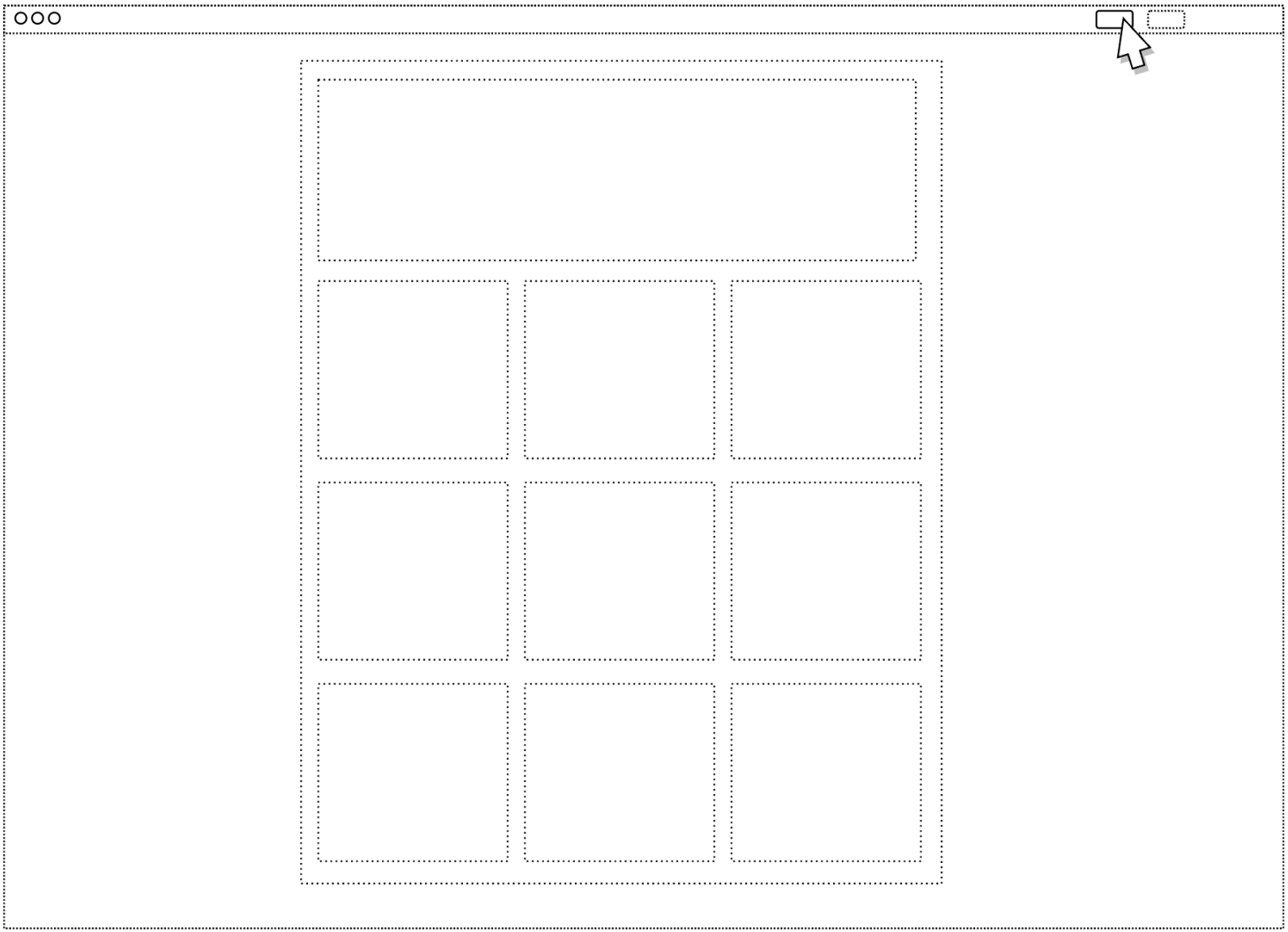
The following pages include summative user flows for the LOG and the STAGE:

A1 LOG
A1.1 ambient CAPTCHA collection
A1.2 manual CAPTCHA collection
A1.3 viewing CAPTCHA collection

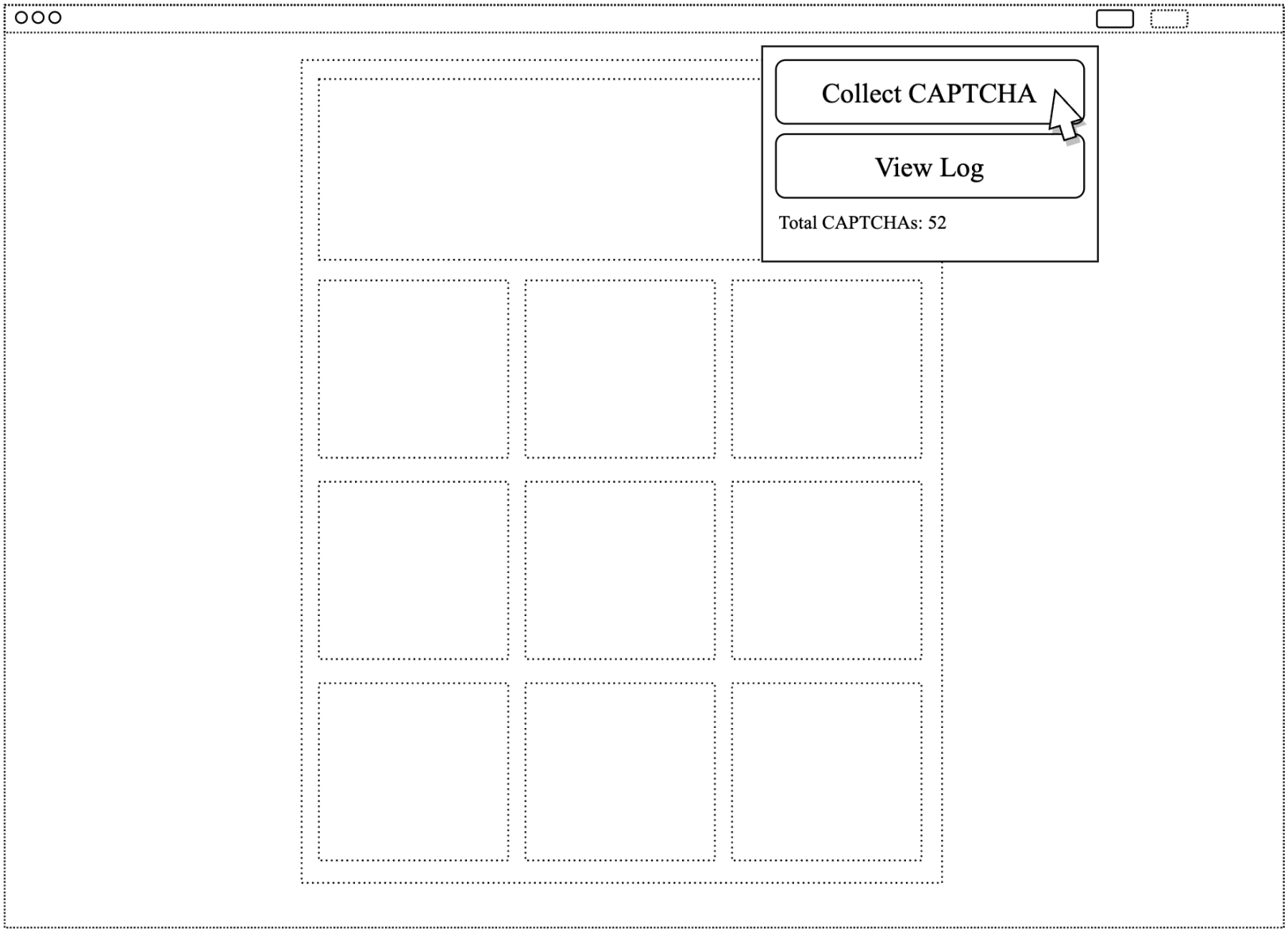
A2 STAGE
a sample user interaction flow:
examining a model laugh, uploading a custom laugh,
and tuning it to fit obscure metrics,
which are persistently displayed.



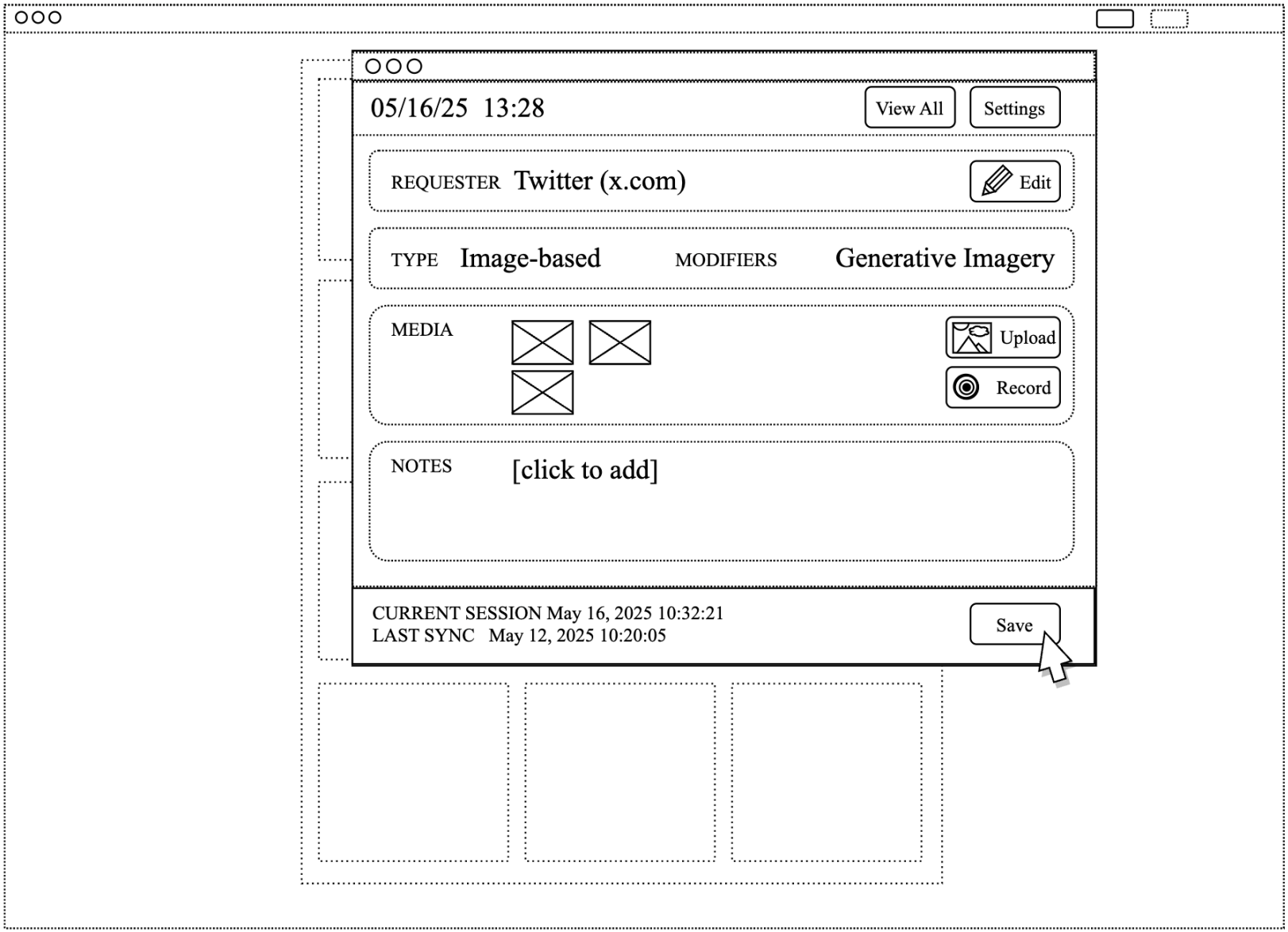
A1.1 LOG: ambient, automatic CAPTCHA collection



A1.2 LOG: manual CAPTCHA collection

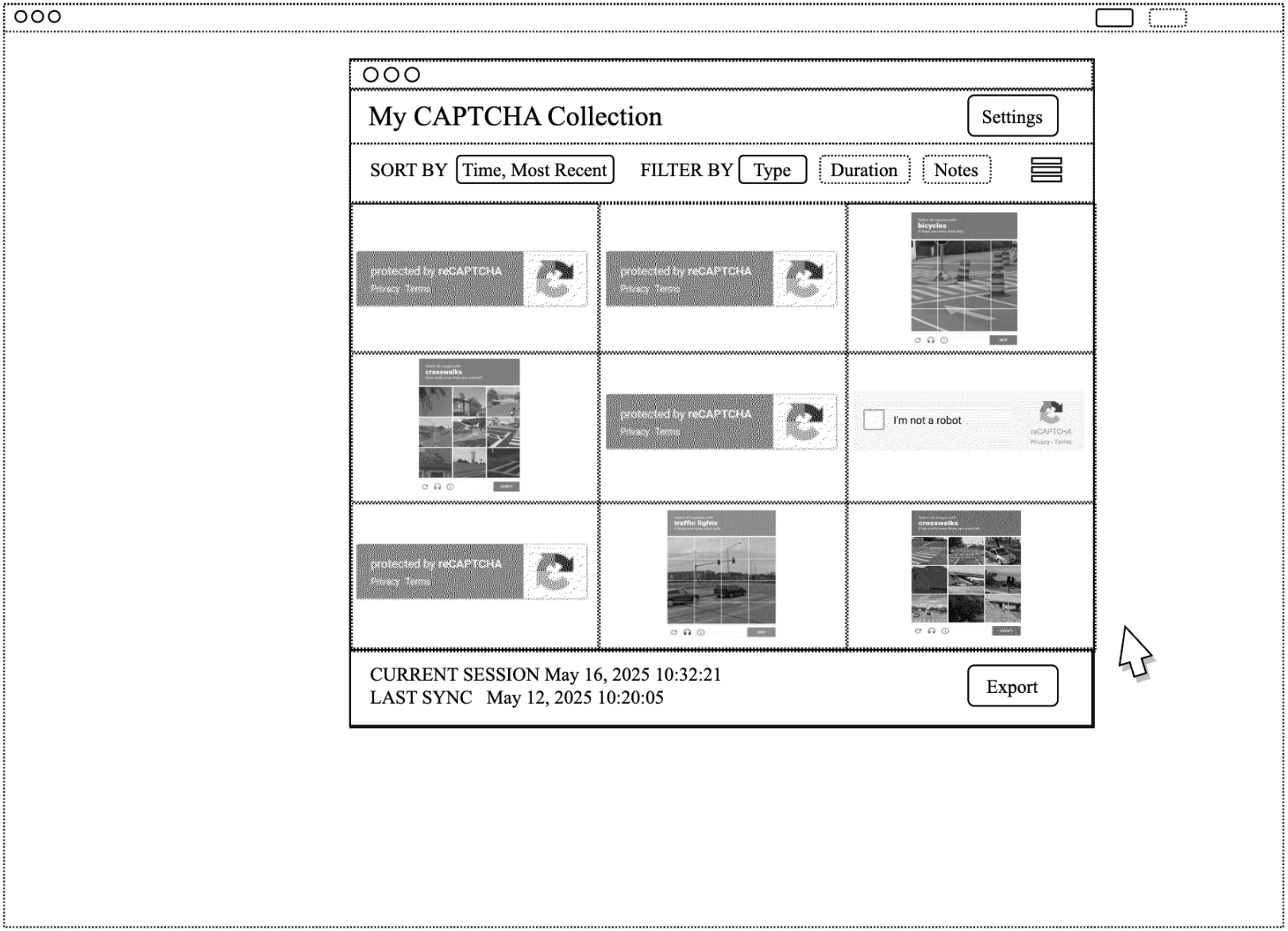


A1.2 LOG: manual CAPTCHA collection

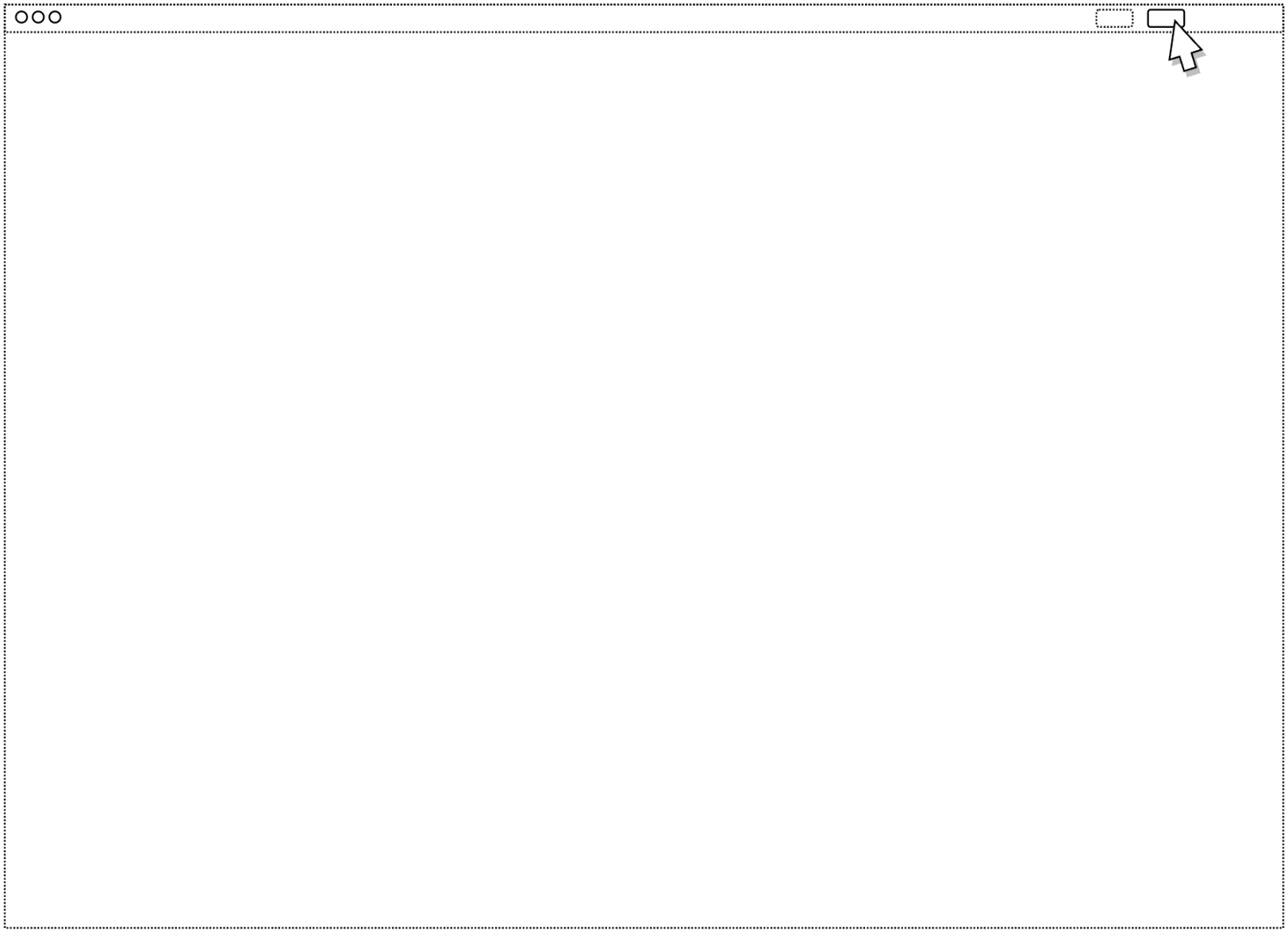


A1.2 LOG: manual CAPTCHA collection

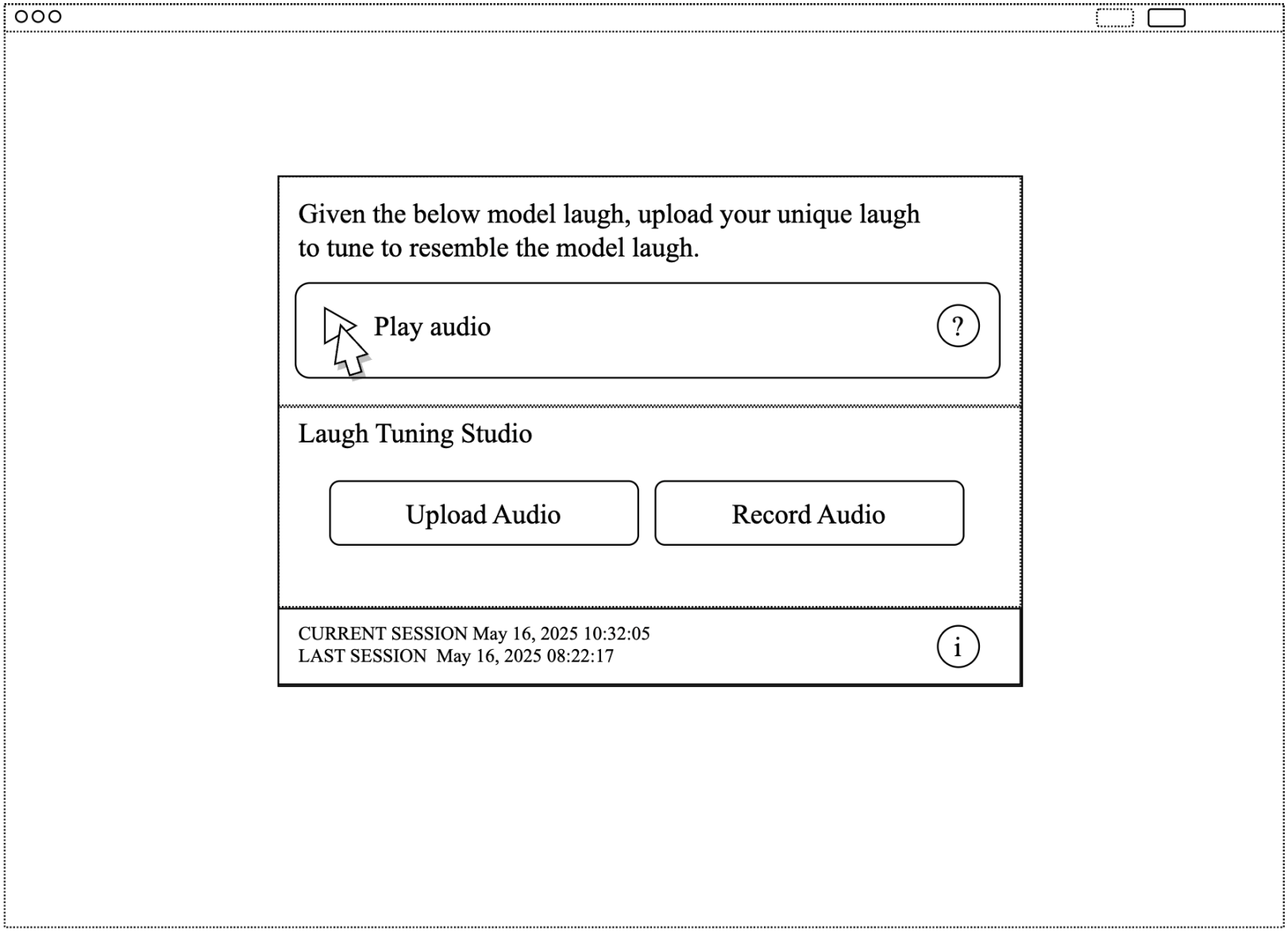
A1.3 LOG: viewing all collected CAPTCHAs



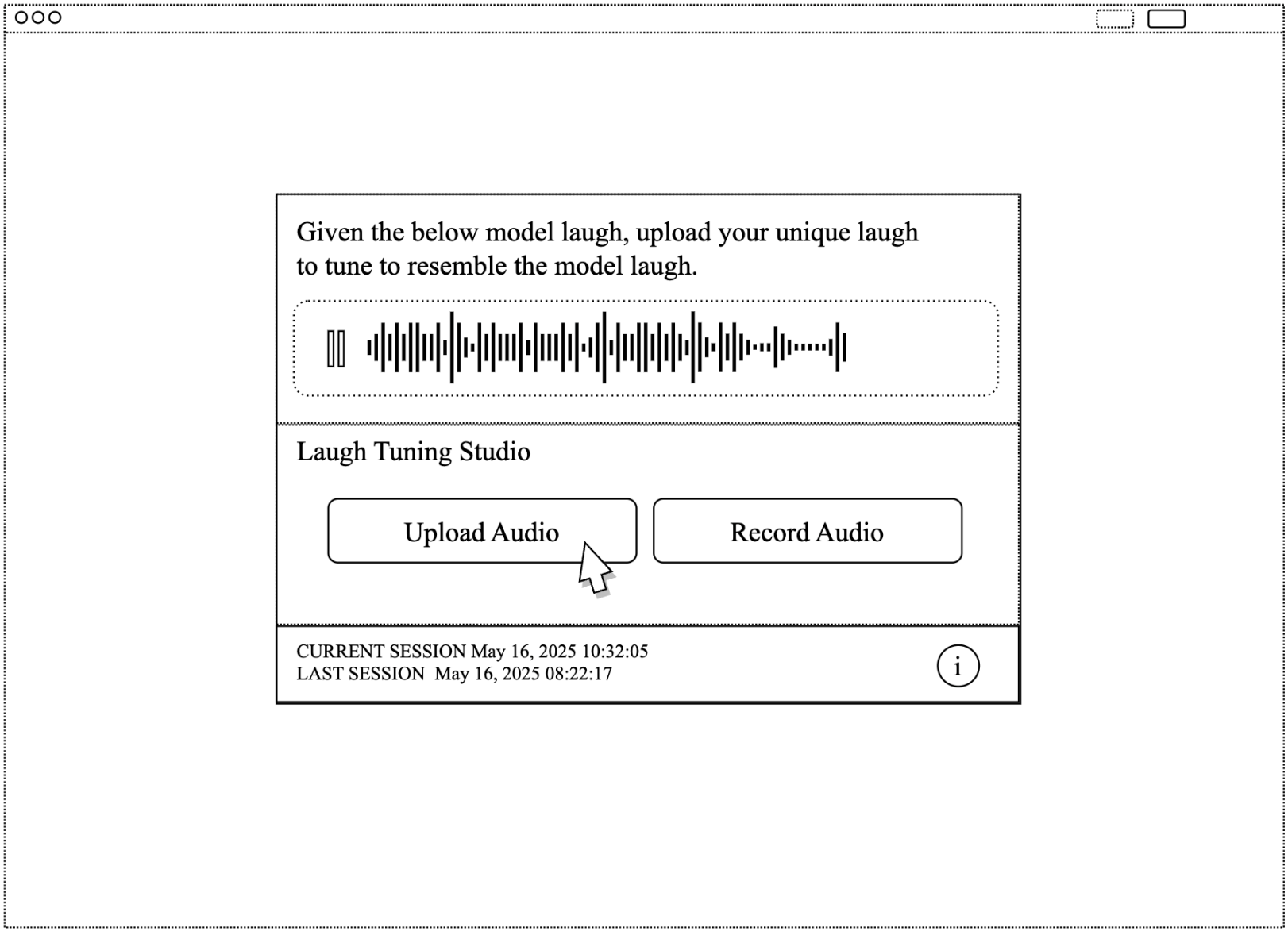
A1.3 LOG: viewing all collected CAPTCHAs



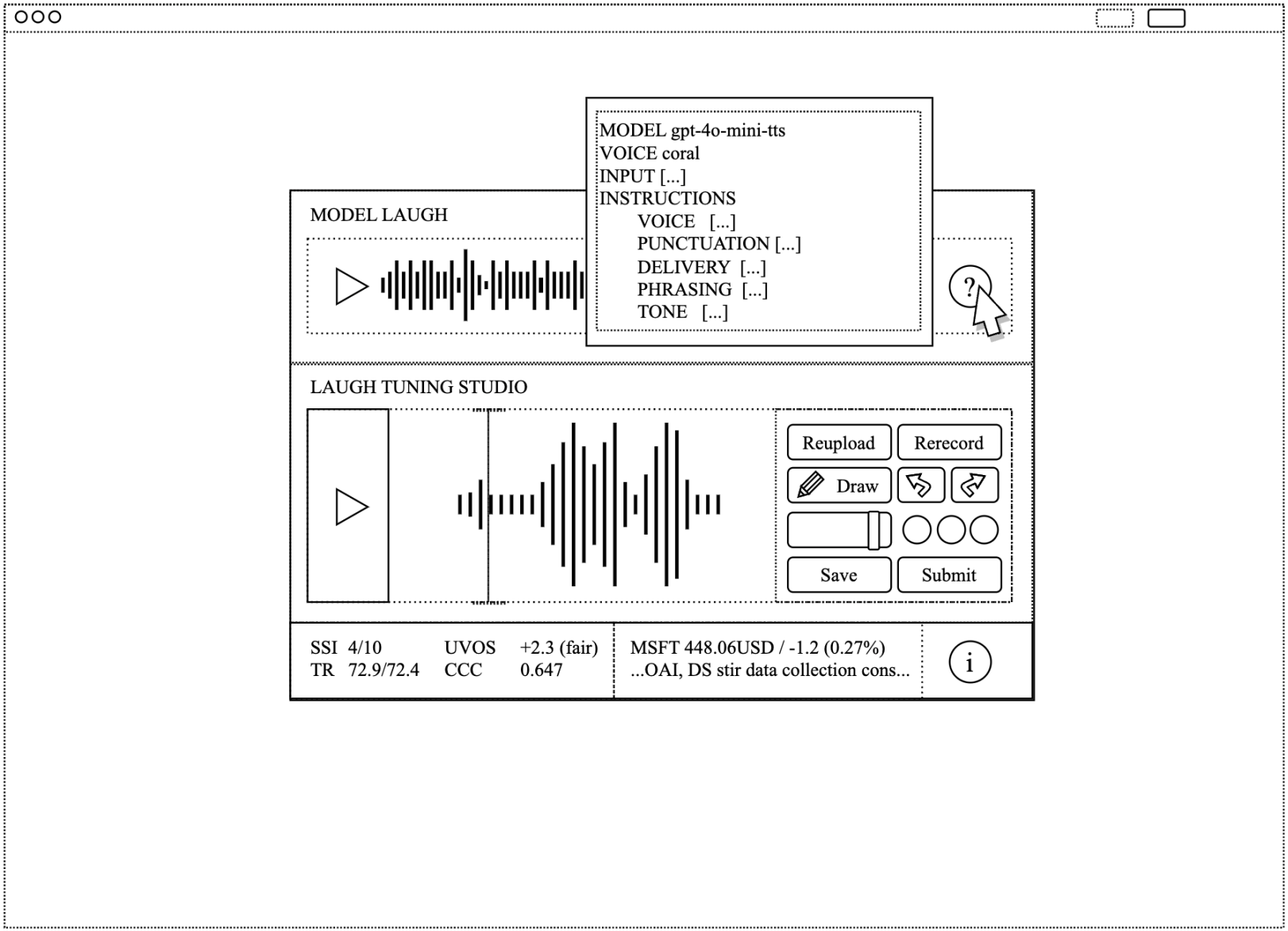
A2 STAGE: tuning a laugh



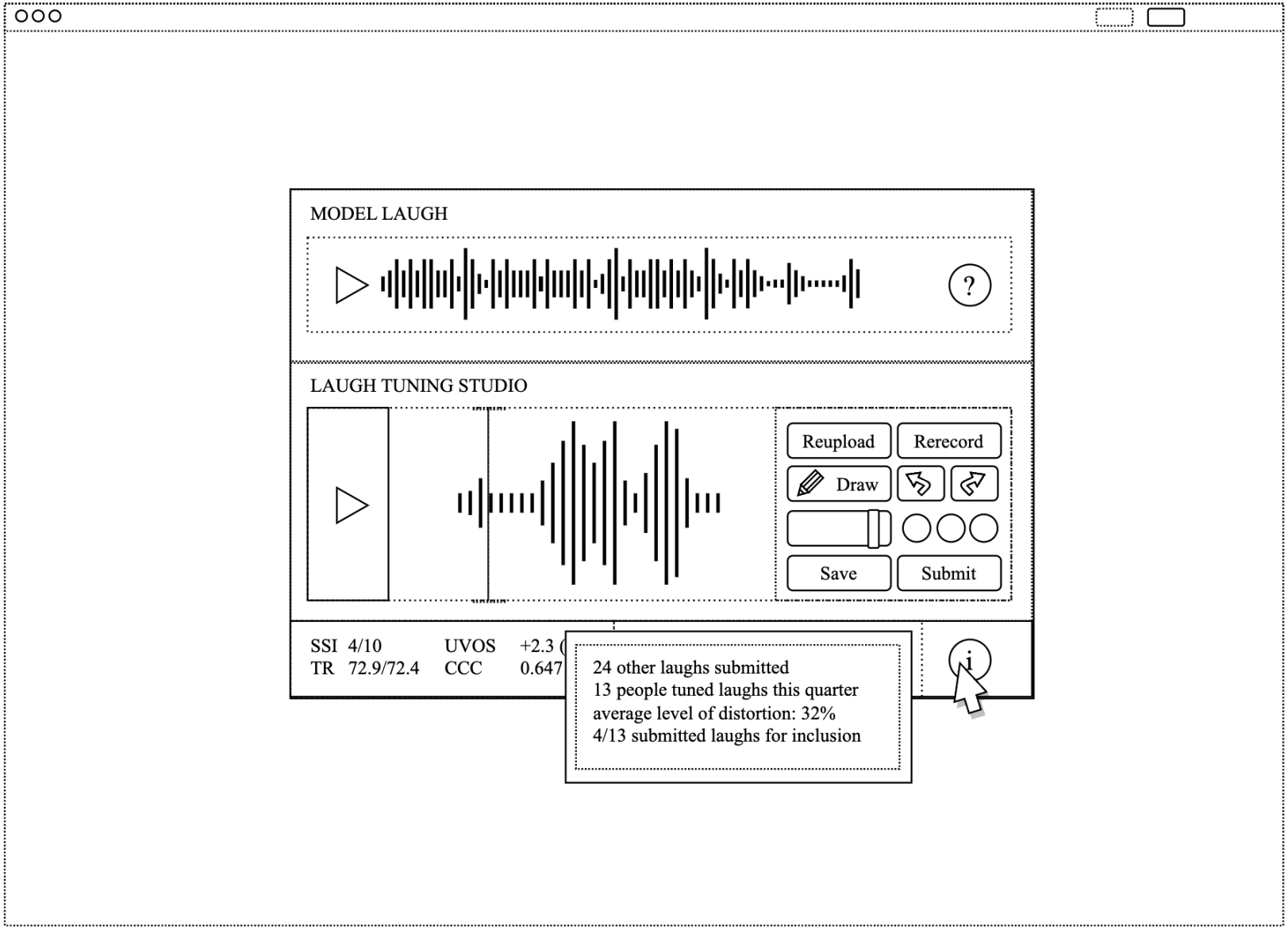
A2 STAGE: tuning a laugh



A2 STAGE: tuning a laugh



A2 STAGE: tuning a laugh
an info panel displays parameters for model laugh (“gold standard”) generation



A2 STAGE: tuning a laugh
additional information about other users situates the activity