PolicyPad: Collaborative Prototyping of LLM Policies

K. J. Kevin Feng
University of Washington
Seattle, WA, USA
kjfeng@uw.edu

Tzu-Sheng Kuo
Carnegie Mellon University
Pittsburgh, PA, USA
tzushenk@cs.cmu.edu

Quan Ze Chen
University of Washington
Seattle, WA, USA
cqz@cs.uw.edu

Inyoung Cheong
Princeton University
Princeton, NJ, USA
iycheong@princeton.edu

Kenneth Holstein
Carnegie Mellon University
Pittsburgh, PA, USA
kjholste@cs.cmu.edu

Amy X. Zhang
University of Washington
Seattle, WA, USA
axz@cs.uw.edu

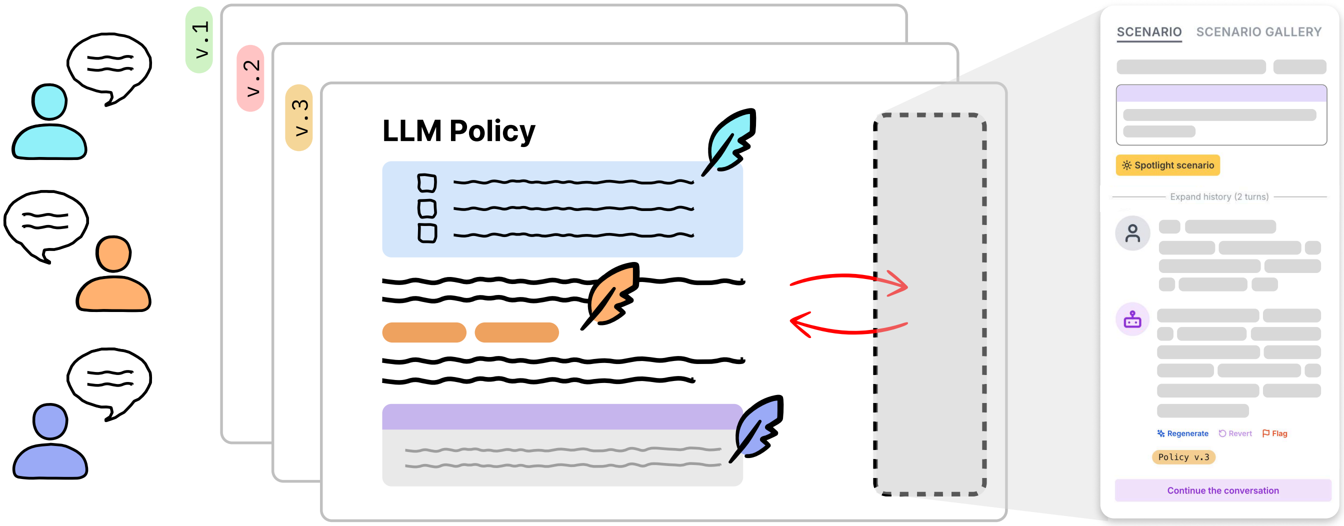


Figure 1: PolicyPad is an interactive system that facilitates collaborative prototyping of LLM policies. Policy designers work together in real time (left) to draft policy statements in PolicyPad's collaborative editor (middle), while experimenting with the model's policy-informed behavior in a private sidebar (right). Content from the private sidebar can be fluidly brought into the collaborative editor for viewing, editing, and discussion. To facilitate LLM policy prototyping, PolicyPad borrows concepts and practices from UX prototyping, including heuristic evaluation, storyboarding, and rapid iteration.

ABSTRACT

As LLMs gain adoption in high-stakes domains like mental health, domain experts are increasingly consulted to provide input into policies governing their behavior. From an observation of 19 policymaking workshops with 9 experts over 15 weeks, we identified opportunities to better support rapid experimentation, feedback, and iteration for collaborative policy design processes. We present PolicyPad, an interactive system that facilitates the emerging practice of *LLM policy prototyping* by drawing from established UX prototyping practices, including heuristic evaluation and storyboarding. Using PolicyPad, policy designers can collaborate on drafting a policy in real time while independently testing policy-informed model behavior with usage scenarios. We evaluate PolicyPad through workshops with 8 groups of 22 domain experts in mental health and law, finding that PolicyPad enhanced collaborative dynamics during policy design, enabled tight feedback loops, and led to novel policy contributions. Overall, our work paves participatory paths for advancing AI alignment and safety.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **Synchronous editors**.

KEYWORDS

LLM policy design, AI alignment, human-centered AI

1 INTRODUCTION

For decades, researchers and science fiction writers have imagined a world in which AI systems can be governed by natural language rules [14, 142]. Today, governing large language models (LLMs) with *LLM policies*—sets of rules, guidelines, and desiderata that shape model behavior—is a key component in the broader toolkit of approaches to improve model alignment and safety [7, 60, 63, 84, 106, 107]. For example, OpenAI's Model Spec contains a series of general objectives and principles (e.g., *"Seek the truth together"*) for researchers and red-teamers to use as a guide when working on reinforcement learning from human feedback (RLHF) [107], as well as for the model to learn from directly [50]. Similarly, Anthropic uses Constitutional AI—in which reinforcement learning receives reward signals from AI-generated feedback that adheres to a set of principles (a "constitution")—to align its Claude models [7, 15]. If effective, LLM policies promise a transparent, familiar, and legible means by which developers and policymakers can govern AI systems.

As LLMs are deployed to millions of users globally, LLM policies become increasingly consequential and scrutinized. This is especially true in high-stakes, tightly regulated domains that are seeing rapid increases in LLM use by everyday users, such as mental health and law [29, 55–57, 93, 117, 127]. LLM policies are primarily written by model developers, but the lack of external input often leads to insular policies that poorly reflect the attitudes and values of everyday users, while ignoring key safety concerns [60, 63]. While frontier AI labs regularly partner with external domain experts to conduct pre-release safety testing of their models [10, 48, 65], there has been little documentation of similar efforts for LLM policies. Yet, there is mounting recognition that expert input is essential for LLM policy design, especially in safety-critical, domain-specific use cases [108, 146].

In this work, we first conduct a 15-week observational study in partnership with a U.S.-based frontier AI lab to better understand how domain experts can contribute to LLM policy design. Through 19 interactive workshops in which mental health experts discussed, annotated, taxonomized, and drafted user queries and LLM responses, we observed experts collaboratively ideating and discussing policy ideas while seeking ways to rapidly test and iterate on them through experimentation with model behavior on realistic scenarios. Much like prototyping in user experience (UX) practice, there is a strong emphasis on collaborative and rapid exploration, feedback collection, and iteration. We thus conceptualize this emerging practice as *LLM policy prototyping*, borrowing from established UX practices like heuristic evaluation and low-fidelity prototyping.

However, few tools exist for policy design [84], let alone tools that support collaborative LLM policy prototyping. We observe notable opportunities for experts to tighten the feedback loop during policy design while leveraging collaborative affordances to build off each other's expertise. We design and develop PolicyPad,[1] an interactive system that facilitates LLM policy prototyping. PolicyPad draws upon established methods and concepts within UX prototyping to enable small groups to collaboratively draft policies, test policy-informed model behavior against usage scenarios, evaluate the quality of the policy, and iterate on its contents in real time through tight feedback loops.

We evaluate PolicyPad through policy prototyping sessions with 22 domain experts spanning two domains—mental health and law—organized into 8 groups. We found that design decisions in PolicyPad fostered collaborative dynamics between experts via its interactive in-editor widgets and yielded more novel policies compared to a baseline, relative to existing policies[2] including OpenAI's Model Spec [107] and Claude's Constitution [7]. Key areas of novelty include offering more specific guidelines on when the model should defer to a human expert, and eliciting key information required for responsible assistance early in the conversation. We end by discussing the practical implications of our work, including where LLM policy prototyping can be situated in AI alignment pipelines, and approaches for scaling up policy prototyping efforts.

Concretely, this work makes the following contributions:

- An 15-week observational study with 9 mental health experts that surfaced opportunities for tight, collaborative feedback loops in LLM policy design.
- LLM policy prototyping, a conceptualization of an emerging practice for collaboratively prototyping LLM policies in small groups.
- PolicyPad, a system that facilitates LLM policy prototyping through interactive and collaborative affordances for policy design, drawing from established UX practices.
- An evaluation of PolicyPad with 22 domain experts in 2 domains, where we found that the system enriched collaboration during policy prototyping and resulted in more novel policies compared to a baseline.

2 RELATED WORK

2.1 Participatory and Collaborative Policy Design

Policies are instruments of governance to guide decisions and establish desirable goals and outcomes [109]. Research has shown that a lack of input from broader constituents has led to governance policies that are poorly informed, fail to address concerns of those affected, and lack legitimacy and trust in the public eye [22, 88, 102]. Thus, expanding the policy-making processes to a wider range of stakeholders through the participatory design of policy can confer many benefits, such as achieving more democratic legitimacy [111], providing better outputs through the integration of localized or specialized knowledge [86], and building social cohesion [100]. Participatory policy design can be especially crucial for public-facing AI and LLM systems [31, 137], as these systems impact large portions of the population [30, 70, 97], while also having many domain-specific risks due to their broad utility [19, 54, 55].

Participatory policy design has been practiced in *offline* governance, with structured processes deployed to create public policies around transportation, environment, technology, and public health around the world [32, 115]. Expert workshops have been one common way to achieve participatory inputs in offline settings, with prominent applications in business [58], public health [130], and politics [4]. *Deliberative democracy* takes this a step further, where representative groups of community members not only contribute knowledge but also deliberate over policy issues [37, 51, 131]. In particular, citizens' assemblies—where small samples of the population participate in workshops where they receive briefings about an issue before deliberating on policy options—have proven to be effective in producing agreeable and actionable policy reports on complex topics while still engaging with largely laypeople constituents [45, 86, 89, 122, 133].

There have been many digital tools developed to augment, scale up, or improve on certain aspects of participatory policy design [47]. Some efforts address scaling by aggregating and distilling opinions

[1] We plan to open-source PolicyPad at https://github.com/kjfeng/policypad.
[2] As of September 1, 2025.

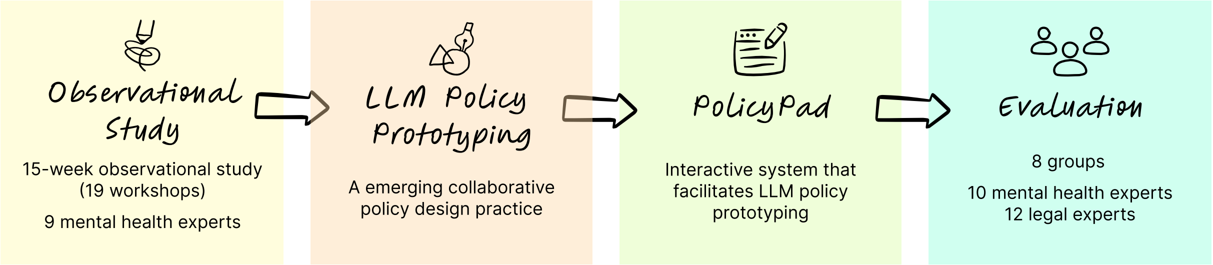

Observational Study
15-week observational study (19 workshops)
9 mental health experts
LLM Policy Prototyping
A emerging collaborative policy design practice
PolicyPad
Interactive system that facilitates LLM policy prototyping
Evaluation
8 groups
10 mental health experts
12 legal experts

Figure 2: Research Process Overview. Our work proceeded in 4 phases: (1) a 15-week observational study with 9 mental health experts (19 workshops) led to (2) conceptualization of LLM policy prototyping. We then (3) designed and built PolicyPad and (4) evaluated it through 8 policy prototyping sessions with 22 experts (10 mental health, 12 legal).

from lower-level preference inputs [8, 15, 105]. Moving beyond simple opinion aggregation, tools like Pol.is [132], Talk to the City [2], Remesh [78, 124], and ConsiderIt [80] also include mechanisms for consensus-finding and balancing of trade-offs. Various tools have also been developed to assist and engage online communities in authoring policies in text or code [83, 96, 150] with others supporting decision-making processes with policies or policy-like guidance [28, 38, 79].

Participatory processes with stakeholder inputs have more recently also been applied to the governance of technology and online platforms. For example, platforms like Facebook have delegated some aspects of content moderation policy to panels with external stakeholders such as the Oversight Board [76, 119]. Other online platforms take a decentralized approach, enabling communities to create and enforce their own policies [67]. Government adoption of AI systems has also faced considerable public deliberation and debate in the US [13, 147] and EU [143]. Increasingly, such policy processes are adopting tools and approaches from participatory policy design, including public surveys and participatory expert workshops in the UK [98] and a citizens' assembly on AI in the EU [49].

Current practical efforts around designing policies for governing LLM behavior have largely been directed by actors within the AI industry. In an effort to ensure the robustness of LLM policies, AI developers and operators often employ in-house teams to surface policy weaknesses through "red-teaming" efforts [1, 39], though many external users and testers may also voluntarily contribute to identifying and reporting such issues as well [149]. More recently, there have also been pushes to formalize safety policies around AI systems via compliance testing organizations [139] and from the academic community [3, 112]. Organizations have also involved the general public through programs aimed at collecting public input on specific topics of policy concern [8, 105, 106]. However, compared to offline participatory efforts, a notably missing component is better integration of *domain experts*—practitioners who have specialized expertise and field experience, but not technical expertise on AI. Compared to laypeople, these domain experts can often give more nuanced input for safety-critical use cases such as mental health, and are thus promising collaborators for designing LLM policies [29]. We thus contribute tooling to support this collaboration in our work.

2.2 Theory and Practice of Prototyping

Prototypes are fundamental artifacts in the design of physical and digital products. At their core, prototypes are representations of an interactive system that simulates one or more aspects of its design and functionality [16, 44, 62, 66, 87, 92]. Depending on what a designer hopes to learn at a particular point in the design process, the prototypes they employ may vary in fidelity, interactivity, duration of use, or representation [16, 62, 129]. For example, prototypes may represent the look and feel of a product, the product's role in users' lives, implementation approaches, or some combination thereof [62]. Lim et al.posit that the most effective prototype is *economical*, one that *"makes the possibilities and limitations of a design idea visible and measurable [...] in the simplest and most efficient way"* [92].

Successful prototypes not only help individual designers explore, conceptualize, and validate design ideas, but also act as boundary objects [134] that facilitate dialogue and collaboration among groups of stakeholders involved in the design process [23, 31, 40, 41, 91, 113, 136]. Specifically, the production of tentative designs that can be quickly iterated upon is an essential component in aligning ideas and expectations in participatory processes [31, 95]. The fidelity of prototypes also significantly impacts their communicative role. Low-fidelity prototypes, such as paper sketches and wireframes, support rapid and parallel ideation by teams early in the design process, while the interactivity and details in high-fidelity prototypes yield more granular feedback to help teams make their ideas more concrete [35, 129]. However, researchers and practitioners have also argued against strict adherence to a low-to-high fidelity prototyping workflow. For example, Virzi et al. [144] found that low-fidelity prototypes can provide effective feedback to designers throughout the product design lifecycle, not just in the early stages. The fidelity level at which prototyping happens should thus be calibrated based on the designers' goals and problems faced [16, 62, 92].

In recent years, researchers and practitioners have started to draw from prototyping practices to improve policymaking processes [52, 74]. *Policy prototyping* has been proposed to enable policymakers to iteratively collect and incorporate feedback on early policy drafts [77, 120, 123]. However, testing and evaluating prototypes of traditional policies often remain intractable in practice due to the need for deployment on real populations and long evaluation timelines. Given that LLM behaviors can be quickly adapted and tested within tight feedback loops for proof-of-concept experimentation and evaluation [9, 40], we observe that *LLM policies* are an especially promising candidate for policy prototyping.

2.3 Methods for Prompt Engineering and Red Teaming

At first glance, LLM policy design shares a key goal with the popular practice of *prompt engineering*: to produce a natural language artifact that shapes model behavior [107, 148]. However, the two have fundamentally different epistemic goals.

In prompt engineering, the goal is **output-driven**—to create a prompt that elicits outputs from a model that satisfy certain criteria [12, 68, 101, 135]. The design of interactive tools for prompt engineering reflects this goal. Tools like Chainforge [12], EvalLM [73], and CoPrompter [69] all allow users to set customized evaluation criteria to compare and evaluate different prompt versions. Because the consumer of the final artifact is the model, factors like legibility and clarity to humans are secondary [11]; the best prompt is one that produces model outputs that best satisfy evaluation criteria [12, 69, 73, 101].

On the other hand, the goal of LLM policy design in our work is **input-driven**—to elicit and encode perspectives on responsible model behavior into a concrete artifact that informs a set of practices for model alignment. The primary consumers of the policy are humans, although the policy may sometimes be used to instruct models too [50]. As a result, factors like legibility and clarity are crucial. LLM policy design may benefit from some prompt engineering practices, such as immediately updating model behavior for quick testing and iteration, applying updates across a range of test cases for evaluation, and version tracking [12, 69, 73]. We thus draw design inspiration from some prompt engineering tools in our work. However, we also design for noteworthy differences, such as scaffolding evaluation of the *policy* rather than *model outputs*.

Related to prompt engineering is the practice of *red teaming LLMs*, where prompts—often adversarial—are perturbed and sent to the model at scale to evaluate the robustness of model safeguards [26, 34, 46, 116]. Like human-centered design, red teaming often benefits from exploring diverse user scenarios to envision edge cases in model use to surface undesirable behaviors [33, 46]. However, the nature of these scenarios varies significantly across the two practices. Red teaming prioritizes scenario quantity—even minor perturbations may be useful in surfacing jailbreak vulnerabilities in models [116]. On the other hand, scenarios in human-centered design serve as provocations for designers and thus prioritize quality and depth [18, 25, 43, 59]. In our work, while some LLM policy design goals are similar to those of red teaming, we focus on supporting in-depth, small-group collaboration, with an aim to surface *both undesirable and desirable model behaviors*, rather than large-scale efforts specifically targeting undesirable behaviors.

2.4 Tools for LLM Policy Design

Due to the nascency of LLM policy design, there are currently few tools to support policy designers, despite the rising importance of LLM policies [63, 84]. Policy Projector by Lam et. al [84] supports AI safety practitioners in authoring if-then rules for LLM content moderation by identifying gaps in existing policies through a visual interface. Our work differs in two important ways. First, rather than patching gaps in existing policies, we support the creation of novel policies that can extend beyond what existing policies cover. Second, we specifically support collaborative policy design, which Lam et al. identified as a fruitful area of future work. ConstitutionMaker [118] allows users to turn written critiques of model responses into principles that guide future behavior, but is a tool for LLM personalization rather than policy design. Roleplay-doh [94] similarly converts written feedback on LLM behavior into principles, but specifically for domain experts to govern LLM-prompted roleplay. We see this feedback-to-principle interaction as valuable to policy designers as well, and integrate a variant of it into our system.

3 OBSERVATIONAL STUDY

3.1 Study Motivation and Procedure

To develop an understanding of real-world LLM policy design practices, we conducted a 15-week-long observational study via contextual inquiry [17] in partnership with a frontier AI lab based in the United States. We wanted to observe how domain experts collaborated to draft domain-specific LLM policies, and any opportunities for improving processes and/or tooling. The lab was convening weekly/twice-a-week virtual workshops (19 workshops total) with 9 experts in clinical mental health (denoted E1–E9) to design new AI policies for model behavior when responding to users' mental health queries. While all 9 experts were invited to every workshop, there were not 9 attendees every week due to scheduling conflicts. Out of the experts, 6 identified as female and 3 as male. For their highest degrees, 4 held a Ph.D. in clinical psychology, 4 held a Doctor of Clinical Psychology (Psy.D.), and 1 held a Master's in Clinical Psychology. Experts were all based in the United States. This study was classified as exempt by our institution's IRB.

At least one member of the research team attended these workshops from January to April 2024. Each workshop was 60–90 minutes in length. One facilitator from either the AI lab or our research team led the workshop with a collaborative policy design activity for the group of experts. These activities revolved around two goals. First, experts developed taxonomies for collections of example mental health-related user queries to an LLM ("scenarios"). Some specific tasks for this goal included deliberating with other experts on taxonomy labels, when to combine and separate labels, and assigning labels to scenarios. Second, experts drafted and voted on desirable rules the model should follow. This included tasks such as reviewing proposed rules in a shared spreadsheet, suggesting modifications, and merging similar rules.

All workshops were recorded and transcribed. As is common in contextual inquiry, team members took notes on observations and asked questions as needed [17]. The first author then deductively coded workshop transcripts based on themes identified in our team's notes, clarifying and iterating on the themes while doing so. We concluded our observational study with a 30-minute semi-structured interview with all the experts in the last workshop. We asked experts to reflect on the workshops and their overall contributions to the policy. The first author used a hybrid inductive-deductive coding process [42] to code the interview transcript. This hybrid process allowed us to connect to themes from our workshop data while embracing new themes emerging from the interview. Our final set of themes can be found in Table 2 in Appendix C.

3.2 Observational Study Results

We used our notes and themes to synthesize four main observations, which we describe in detail below.

3.2.1 Incomplete feedback loops without model experimentation. Throughout the workshops, experts had visions for how their contributions to the policy—through drafting taxonomies and principles—can impact model behavior. For example, E3 shared that *"a clinical minimization [of the user's feelings] can be helpful, but for the model, that would be hard to decipher"* so the group wrote a policy barring the model from engaging in this behavior. However, experts failed to verify whether and how those visions *actually came into fruition* because they did not have a "policy-informed" model—a model that acts in accordance with the policy—to interact with. This resulted in an incomplete feedback loop. Experts were unable to obtain signals about the effectiveness of their policy contributions and any unintended side effects that may arise, as E9 explains: *"Just because you think that might be a good rule, it may have an unanticipated consequence you don't realize. I think that it would be really helpful to know how these [rules] we're coming up with actually play out."* E7 agreed and added that direct experimentation with a policy-informed model would allow them to better *"see how [a conversational interaction] would play out from the perspective of a user."* Experts had unrestricted internet access throughout the workshops and could test behaviors out on popular chatbots. However, we did not observe instances of this, possibly due to preoccupation with workshop activities, or lack of knowledge about (or in some systems, inability to set) custom system prompts.

3.2.2 Experts tackled both high-level strategy and low-level semantics. We noticed that experts could easily derail from workflows that would enable them to best contribute their expertise when designing policies. For example, when creating taxonomies for mental health-related user queries, experts spent substantial time wrestling with wording and semantics. Similarly, E9 reflected that much of their time was spent on finding the right wording for taxonomy labels: *"we thought needed to not spend forever trying to wordsmith exactly how that needed to appear."* In a separate activity where experts wrote out ideal model responses, E5 agreed that experts should avoid getting stuck in the weeds of low-level wording edits: *"It would be more effective at this stage for us to just put our thoughts in about what's right or wrong, because the time it takes to craft the perfect response is out of scope for this task."* Study facilitators agreed that much of the low-level semantics can be refined post-hoc via LLMs, as long as there are sufficient amounts of expert insight to guide that refinement.

3.2.3 Scenarios grounded discussions and spurred policy generation. We found that experts engaged in richer discussions that led to insightful policy suggestions after they were given scenarios—examples of user-AI conversations that may arise in real-world use—for reference. For E1, looking at scenarios helped them identify two pieces of information the model should consider in its response: *"We need to ask clarifying questions, in particular to clarify the severity and the nature of the dark thoughts this person suggested. Another dimension is to identify how long they've been feeling this way and what sources of support they have."* E2 agreed with the need for a severity assessment, suggesting a safety rating scale for the user in case they cannot quickly reach a professional and need an immediate response: *"The AI needs to respond, providing resources quickly. Maybe having a rating scale on the scale of zero to 10, how safe are you feeling right now?"* Adding on, E3 suggested eliciting the user's financial ability to pay for therapy and making referrals accordingly: *"There might be questions instead like, what is your financial ability to pay for therapy right now? And if it's within certain ranges, then you might make a community mental health referral, like here's some Medicare people in your area."* Exploring scenarios helped experts spot recurring problems in model responses and turn them into clear policies. While rules should stay broad enough to be useful, it is unclear how specific they should be. When scenarios show patterns that keep causing issues, they become obvious candidates for new rules. as E7 describes: *"I keep seeing this thing over and over and it's incorrect, so that needs to be a rule."*

3.2.4 Experts valued synchronous collaboration. In contrast with prior work that collected human feedback via asynchronous annotation (e.g., [15, 110]) and/or focused on asynchronous policy design [83, 84], our workshops engaged experts in *synchronous* collaboration—drafting, discussing, and iterating on policy in real time. Experts unanimously agreed that synchronous collaboration was enjoyable and productive. In E1's words, *"I found it hugely rewarding and beneficial personally and professionally [...] I think we can get stuck in our heads because we're working on our own with our clients so much. It was really nice to hear other people's perspectives and thoughts."* E6 emphasized the support and learning opportunities afforded by collaboration: *"[it was] very supportive having other voices in the back of your head [...] it's been incredible learning with everyone."* E9 found synchronous collaboration important for surfacing new perspectives and broadening coverage of the policy: *"[...] there were times where someone else said something that just never occurred to me. We all know one person's opinion is never sufficient, especially in an area as diverse as mental health."* Broadly, we observed that experts were able to quickly resolve disagreements and draft policy statements that had broad consensus in a synchronous setting.

4 LLM POLICY PROTOTYPING

Our observations in Section 3 posed challenges that are not foreign to HCI; well-established concepts and methods in UX design and prototyping can offer help in mitigating these challenges. We now use this insight to conceptualize **LLM policy prototyping** (henceforth "policy prototyping" for brevity), an emerging practice by which groups of individuals can synchronously collaborate on designing an LLM behavioral policy. Specifically, we map observations we identified in Section 3.2 to relevant UX methods, which are then mapped to their usage in policy prototyping. This is depicted in Table 1. We focus our work on *low-fidelity* policy prototypes (row 2 in Table 1)—artifacts with the primary goal of eliciting and integrating group perspectives on responsible model behavior, rather than a high-fidelity, "production-ready" policy. We leave the translation of low- to high-fidelity policies to future work.

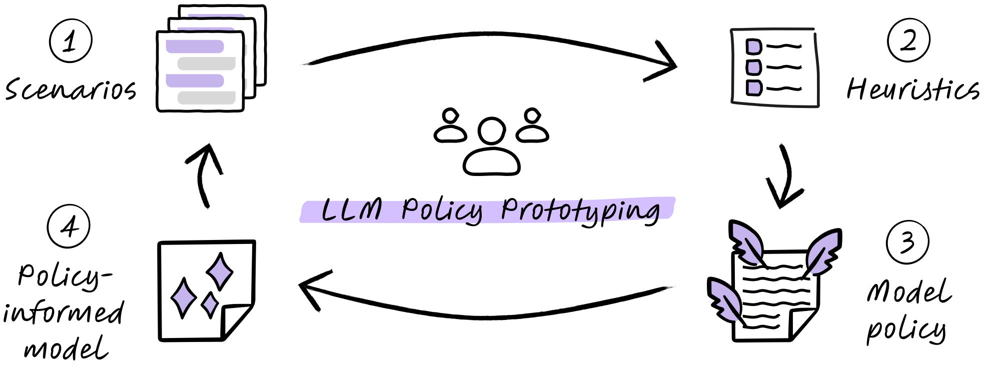


Figure 3: Illustration of our envisioned *LLM policy prototyping* process. Scenarios inform desiderata for the policy via heuristics, which in turn guide the design of the policy. The policy shapes the behavior of a policy-informed LLM, which designers can then test against the scenarios to observe changes in behavior. The process is iterative: feedback from testing may lead to the creation of new scenarios, heuristics, and policy statements.

Concretely, we propose policy prototyping for a policy P in domain D to involve the following activities, as illustrated in Figure 3.

(1) Policy designers review a small set of scenarios that accurately depict real-world use of AI in a specific domain. This allows designers to better understand current AI behavior and contexts of use.
(2) Informed by the scenarios, designers finalize a set of heuristics they will use to ensure P preserves its quality and focus across many iterations.
(3) Guided by the heuristics, designers collaboratively draft policy statements that they believe will lead to more responsible model behavior in D.
(4) Designers test a policy-informed model that acts in accordance with P with existing scenarios. New scenarios, heuristics, and policy statements may be created based on insights from testing and discussions with other designers. The process then repeats until the prototyping session concludes.

5 POLICYPAD

We introduce PolicyPad, an interactive and collaborative system for synchronous LLM policy prototyping. We first describe how we arrived at our final design through three co-design sessions. Then, we walk through the system and its features.

5.1 Iterative Co-Design Sessions with Experts

We designed PolicyPad iteratively through co-design sessions with the same participants in our observational study. These sessions were conducted towards the end of our observational study period. In each session, we presented an interactive prototype of the system.[3] We then collected semi-structured feedback from participants and iterated on the prototype based on feedback for the next session. We repeated this until data saturation—participants were no longer able to provide substantial feedback until we implemented the system, for a total of three sessions (c.f. [82]).

We provide detailed documentation of how we integrated participants' feedback to iterate from a basic first version to our final design in Appendix D. In the following section, we illustrate PolicyPad's capabilities using a system walkthrough.

5.2 System Walkthrough

PolicyPad can be used by any individual or group who wishes to facilitate a policy prototyping session—whether it be an AI lab, academic group, non-profit, or another organization. As a note on this section's terminology, we distinguish "facilitators" (those running the policy prototyping session) from "users" (policy designers participating in the session).

5.2.1 Preliminaries. When users log into PolicyPad, they see a collaborative document editor (Fig 4 **B**), similar to Google Docs. The facilitator may provide light starting materials for the policy, such as high-level objectives, an initial set of policy heuristics (perhaps drawn from trust & safety literature), and a few scenarios for the group to work with. Ideally, scenarios are representative of real-world model use in a domain. For example, facilitators who have access to chatbot logs may use privacy-preserving conversations from their logs.

Users can access these scenarios via the **scenario gallery** in the right sidebar (Fig. 4 **C**). While the document is a collaborative workspace, the sidebar is private to each user, allowing for independent experimentation with the policy-informed model.

5.2.2 Scenario sidebar. A user can browse the scenarios in the gallery and open a scenario in a detailed view (Fig. 4 **D**). The scenario expands to fill the sidebar with the full user-AI conversation, as well as a brief, AI-generated summary of the conversation's contents thus far. As the group reads the scenarios, they start to develop ideas for what to include in the policy.

Formally, a scenario in PolicyPad comprises of three parts: the background (all messages in the conversation up until the most recent turn), the newest user message, and the newest AI message. Given the background and the newest user message, the policy-informed model generates the newest AI message.

5.2.3 Interactive in-editor scenario widgets. Users can bring a scenario from their own scenario sidebar into the collaborative editor

[3] Two prototypes were in Figma. One was implemented in TypeScript and React.

Observation	Relevant UX Method	UX Definition & Usage	Usage in LLM Policy Prototyping	Core Literature
Incomplete feedback loops (Section 3.2.1)	**Rapid Prototyping**	Tight feedback loops of **ideating, implementing, and evaluating design ideas**. Allows designers to identify usability issues early, explore alternatives, and align teams to shared visions	Tight feedback loops of **ideating, drafting, and testing policy statements** for quick identification of policy "usability" issues (e.g., unclear statements), characteristics of responsible model behavior, and translations of that behavior into policy.	[23, 24, 35, 62, 92]
Operating at both high and low levels (Section 3.2.2)	**Low-fidelity prototyping**	Artifact **loosely** resembling the final product in terms of look & feel and/or implementation. Cheap to create and iterate upon, ideal for collecting early requirements and feedback.	Artifact providing **high-level documentation** of responsible model behavior to quickly gather and integrate perspectives/feedback. Requires focus on **high-level** details. **We focus on this type of policy prototype in our work.**	[129, 144, 145]
Operating at both high and low levels (Section 3.2.2)	**High-fidelity prototyping**	Artifact **closely** resembling the final product in terms of look & feel and/or implementation. They are useful for collecting detailed feedback but may be expensive to create.	Artifact providing **detailed documentation** of responsible model behavior to guide alignment efforts, often with polished wording, illustrative examples, legal sign-off, and more. **Requires focus on high- and low-level details.** Example: OpenAI Model Spec [107].	[129, 144, 145]
Scenarios grounded discussions (Section 3.2.3)	**Storyboarding**	Concrete representations of **users, contexts, and tasks** to ground abstract design ideas. **Panels** add context and illustrate user stories. Promotes reflection and communication among stakeholders.	**Sample user-AI conversations** to ground policy discussions and creation. **Conversational turns** add context and illustrate sample user & model behaviors. Promotes reflection and communication among stakeholders.	[4, 18, 43, 59, 141]
Experts valued synchronous collaboration (Section 3.2.4)	**Design workshops**	Common collaborative method for gathering user requirements, studying empirical phenomena, and evaluating interactive systems. Can serve as a field site, research instrument, or a research account.	Small-group sessions that serve as a **field site** for collective ideation and reflection of responsible model behavior in domain-specific use cases.	[36, 126, 128]

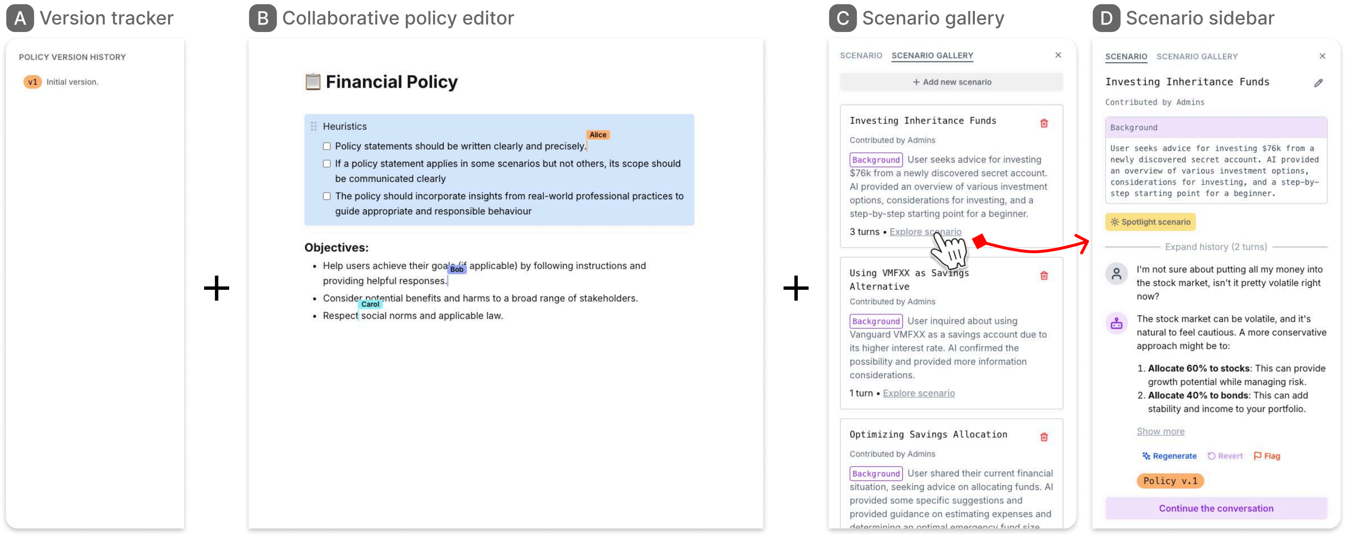
Table 1: Mapping of UX methods relevant to insights from our observational study (Section 3) to their usage in LLM policy prototyping.

by referencing its title with the '@' symbol. Once referenced, the scenario appears inline in the editor as an interactive, pill-shaped widget (Fig. 5). When a user clicks the widget, the full scenario will be shown in their scenario sidebar. These widgets can be used as illustrative examples of model behavior and build shared context when designing the policy. For example, a user may observe that the model does not provide disclosures of capability limits, or incorrectly assumes a detail not explicit in the conversational context. They can flag a model response in their scenario sidebar, which will make the scenario widget glow orange in the editor, encouraging others to take a look.

5.2.4 Drafting policy statements. Once a group reviews the heuristics at the top of the editor to ensure they have a common understanding of policy desiderata, they are ready to start drafting policy statements. These policies address oddities, concerns, and other noteworthy aspects of model behavior they observed and flagged in the scenarios.

As an example, for a policy on providing responsible financial advice, a user may add a policy statement instructing the model to use *cautious, neutral, and non-prescriptive language* while *always surfacing a brief disclosure of limitations early in the conversation.* A couple users may collaboratively draft a policy statement for the model to *defer the user to a licensed adviser or a compliant robo-advice product that meets regulatory obligations.* The group can review the policy together and take advantage of the real-time collaborative editing features to further refine each other's statements.

5.2.5 Testing the policy with scenarios. Users can independently experiment with the behavior of the policy-informed model by *regenerating responses* in the scenario sidebar (Fig. 7 **1**). Independent testing allows each user to focus on the specific concerns that drive their policy contributions, explore challenging boundary cases, and conduct stress-testing without group dynamics influencing their approach. Once they save the policy, they can also browse and compare responses generated by past policy versions.



POLICY VERSION HISTORY

v1 Initial version.

📋 **Financial Policy**

Heuristics

- □ Policy statements should be written clearly and precisely.
- □ If a policy statement applies in some scenarios but not others, its scope should be communicated clearly
- □ The policy should incorporate insights from real-world professional practices to guide appropriate and responsible behaviour

Objectives:

- Help users achieve their goals (if applicable) by following instructions and providing helpful responses.
- Consider potential benefits and harms to a broad range of stakeholders.
- Respect social norms and applicable law.

Investing Inheritance Funds

Contributed by Admins

Background User seeks advice for investing $76k from a newly discovered secret account. AI provided an overview of various investment options, considerations for investing, and a step-by-step starting point for a beginner.

3 turns • Explore scenario

Using VMFXX as Savings Alternative

Contributed by Admins

Background User inquired about using Vanguard VMFXX as a savings account due to its higher interest rate. AI confirmed the possibility and provided more information considerations.

1 turn • Explore scenario

Optimizing Savings Allocation

Contributed by Admins

Background User shared their current financial situation, seeking advice on allocating funds. AI provided some specific suggestions and provided guidance on estimating expenses and determining an optimal emergency fund size

Investing Inheritance Funds

Contributed by Admins

Background

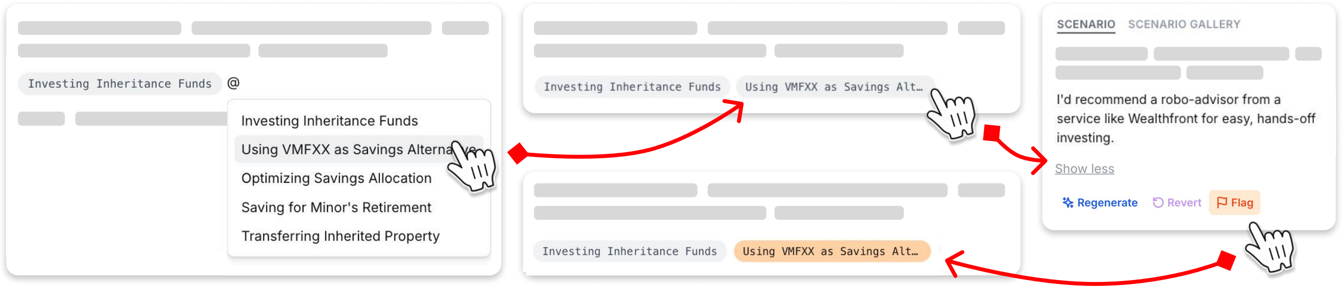
User seeks advice for investing $76k from a newly discovered secret account. AI provided an overview of various investment options, considerations for investing, and a step-by-step starting point for a beginner.

I'm not sure about putting all my money into the stock market, isn't it pretty volatile right now?

The stock market can be volatile, and it's natural to feel cautious. A more conservative approach might be to:

1. **Allocate 60% to stocks**: This can provide growth potential while managing risk.
2. **Allocate 40% to bonds**: This can add stability and income to your portfolio.

Figure 4: Main components of the PolicyPad system. Users can keep track of their policy version in the left sidebar (A) as they collaborative edit the policy in the editor (B). Users can access scenarios via the scenario gallery (C). When they click into a scenario, they can view its full details and explore how the policy-informed model will behave on it via the scenario sidebar (D).


Investing Inheritance Funds
@
Investing Inheritance Funds
Using VMFXX as Savings Alternative
Optimizing Savings Allocation
Saving for Minor's Retirement
Transferring Inherited Property
Investing Inheritance Funds
Using VMFXX as Savings Alt…
SCENARIO
SCENARIO GALLERY

I'd recommend a robo-advisor from a service like Wealthfront for easy, hands-off investing.

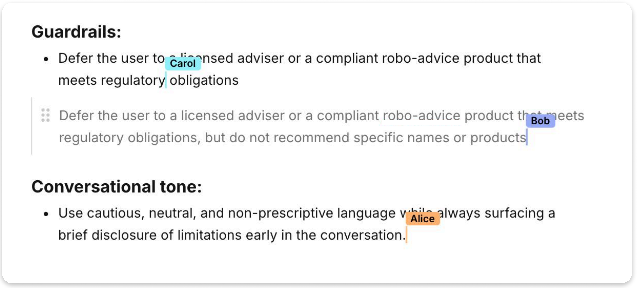
Show less
Regenerate
Revert
Flag
Investing Inheritance Funds
Using VMFXX as Savings Alt…

Figure 5: Scenarios can be brought into the editor inline with the policy as interactive widgets via referencing the scenario's title with the '@' symbol. Once in the editor, all users can click on it, view it in their scenario sidebar, and flag responses for group discussion.

To propose edits in a non-disruptive way, users can add a **drafting block** (Fig. 6) in the editor. Just like how a comment in a code editor is visible to a programmer but does not affect program behavior, content in the drafting block is visible to the group but is ignored by the model. After users review and reach consensus on changes, content in the drafting block can be integrated into the actual policy

To stress-test a policy, a user can *extend* a scenario in the sidebar by continuing the existing conversation (Fig. 7 **2**). This offers an alternative way to experiment with policy-informed model behavior beyond regenerating a single message in a scenario.

Group members are not limited to only the scenarios initially provided to them. They may extend an existing scenario and add the extended version to the scenario gallery for others to view and extend further. They may also create a new scenario from scratch if none of the existing scenarios explore a particular behavior they want to test.



Guardrails:

- Defer the user to a licensed adviser or a compliant robo-advice product that meets regulatory obligations

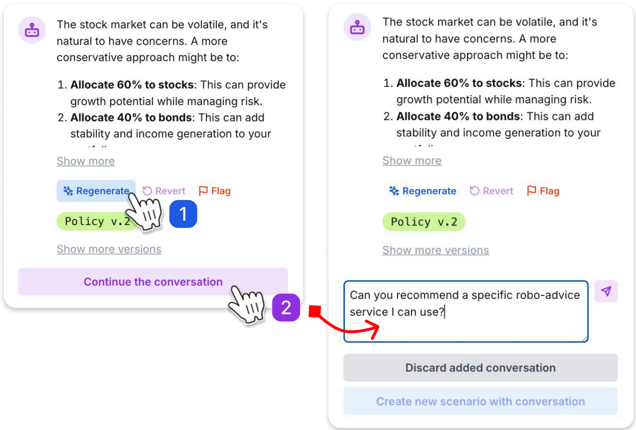
Defer the user to a licensed adviser or a compliant robo-advice product that meets regulatory obligations, but do not recommend specific names or products

Conversational tone:

- Use cautious, neutral, and non-prescriptive language while always surfacing a brief disclosure of limitations early in the conversation.

Figure 6: A drafting block (directly above "Conversational tone") can be added into the editor to draft experimental policies without affecting model behavior.

After a group has made meaningful edits to the policy, they can **save a new version**. After a user clicks the **[Snapshot policy]**



The stock market can be volatile, and it's natural to have concerns. A more conservative approach might be to:

1. **Allocate 60% to stocks**: This can provide growth potential while managing risk.
2. **Allocate 40% to bonds**: This can add stability and income generation to your

The stock market can be volatile, and it's natural to have concerns. A more conservative approach might be to:

1. **Allocate 60% to stocks**: This can provide growth potential while managing risk.
2. **Allocate 40% to bonds**: This can add stability and income generation to your

Can you recommend a specific robo-advice service I can use?

Figure 7: PolicyPad offers two ways to test the behavior of the policy-informed model against a scenario: (1) regenerating the latest AI message, or (2) continuing the conversation.

button, PolicyPad will add the current policy to the version history and regenerate the newest AI message for all scenarios using the policy-informed model (Fig. 8 **1&3**).

5.2.6 Scenario spotlights. Once a user has referenced a scenario in the editor, they can **spotlight** it to expand the interactive widget into a card UI that makes the full scenario visible to everyone (Fig. 9). Once spotlighted, the group can view, discuss, and even collaboratively edit the policy-informed model's response. Just like other text in the document, the scenario spotlight supports real-time collaboration for editing. After users are satisfied with their edits, any user can save the response. The old response remains easily accessible through a simple toggle.

PolicyPad then automatically analyzes the group's edits in the context of the current policy and heuristics. It uses a reasoning LLM to **suggest a policy statement** designed to steer the model towards producing a response more similar to the edited version (Fig. 9 **4**). If the user accepts the suggestion, it will be integrated into the policy. This alternative way of indirectly editing a policy through editing the model response is inspired by prior work on synthesizing principles from edits [94, 118].

Once a group is finished with a scenario spotlight, a user can un-spotlight the scenario for everyone, shrinking the card back into a small, pill-shaped widget.

5.2.7 Heuristics editor and evaluator. As a group expands and refines their policy, they may encounter additional considerations they would like to add as policy heuristics. For example, a policy may become increasingly riddled with domain-specific terms and acronyms unfamiliar to a layperson reading the policy, so the group may add a heuristic for clearly defining or explaining these terms. Since the likely audience of this policy is other people, factors like clarity and legibility are important to preserve.

Every time the policy is saved, PolicyPad runs an automated heuristic evaluation to highlight any unsatisfied heuristics (Fig. 8 **2**). This automated evaluation is meant to draw attention to the heuristics and encourage discussion around them, rather than conclusively determining their fulfillment. Group members can easily override the automated decision if they agree on a different assessment.

Users continue to engage in this iterative process of policy drafting and experimentation until the session concludes.

5.3 Technical Details

PolicyPad is implemented as a web application built with React, TypeScript, and TipTap[4]. The real-time collaboration engine is supported via TipTap Cloud. PolicyPad uses serverless functions to call the OpenAI and Together.ai APIs. The policy-informed model is an instance of Llama 3.3 70B Instruct Turbo[5] hosted on Together.ai. The policy was fed into the model as a system prompt with some additional scaffolding to ensure the model followed it. We called GPT-4o for miscellaneous features that required light LLM processing (e.g., generating titles of new policy versions that capture key changes), and o4-mini for features that benefited from deeper reasoning (i.e. suggesting policy statements after response edits and automated heuristic evaluation).

6 EVALUATION STUDY

To evaluate PolicyPad, we ran a series of group-based, within-subjects studies with 22 domain experts from two domains (10 from mental health, 12 from law). Our goal was to determine how the design decisions made for PolicyPad enhanced the policy prototyping experience. We also evaluate the outputs of PolicyPad by analyzing the policies created by experts to determine their novelty with respect to established LLM policies like Claude's Constitution [7] and OpenAI's Model Spec [107] (as of September 1, 2025). Thus, we asked the following research questions:

RQ1: How did the individual components of policy prototyping supported by PolicyPad (rapid policy iteration, heuristic evaluation, interaction with scenarios, real-time collaboration) aid expert-driven policy design in practice?

RQ2: To what extent are the insights in experts' policies created through PolicyPad novel compared to existing, publicly available LLM policies?

We selected mental health and law as our domains because they are regulated, high-stakes domains for which AI use has been increasing but contested[6] [29, 85, 93, 103]. Crafting responsible LLM policies is therefore critical for ensuring users' safety and well-being. The two domains are also distinct enough for us to observe how approaches to policy prototyping and the resulting policies can differ across domains.

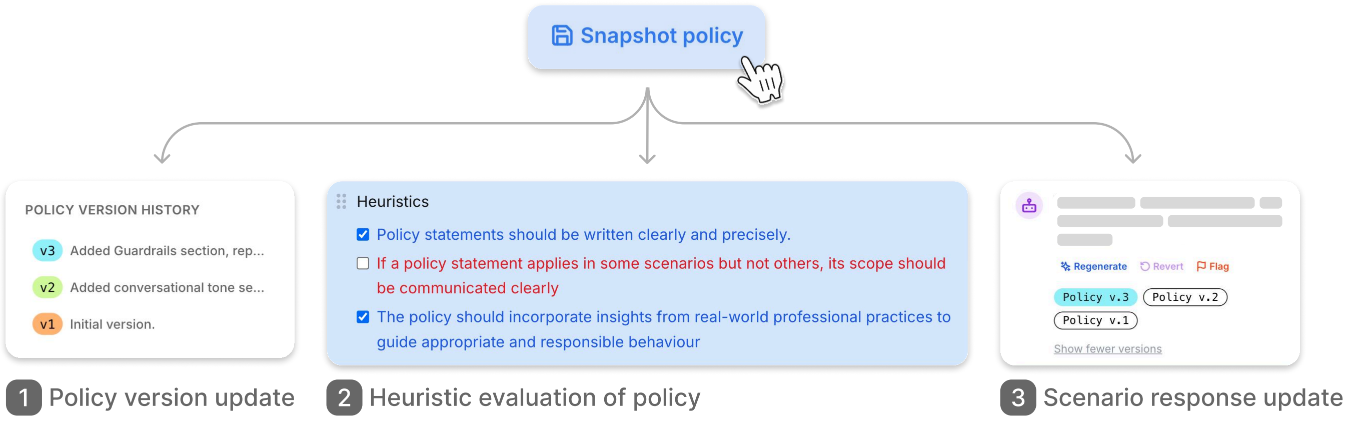
6.1 Participants and Setup

We recruited 22 domain experts (Table 3) through our personal connections, university mailing lists, professional Slack channels, and snowball sampling. Among mental health experts ($n = 10$, 7

[4] https://tiptap.dev/.

[5] The ideal model for policy prototyping generates comparable responses to popular chatbot products (e.g., ChatGPT, Claude) but does not have an existing policy trained into it outside of basic guardrails. Llama 3.3 is an apt fit given its capable performance and bring-your-own safeguards setup (which we did not use). We also used Llama 3.3 instead of Llama 4 due to reports of the latter being narrowly optimized for specific benchmarks [125].

[6] OpenAI CEO Sam Altman observed that the younger generation uses it as a *"therapist, a life coach [...] asking, 'What should I do?'"* despite the absence of confidentiality protections that govern human attorneys and therapists [117].

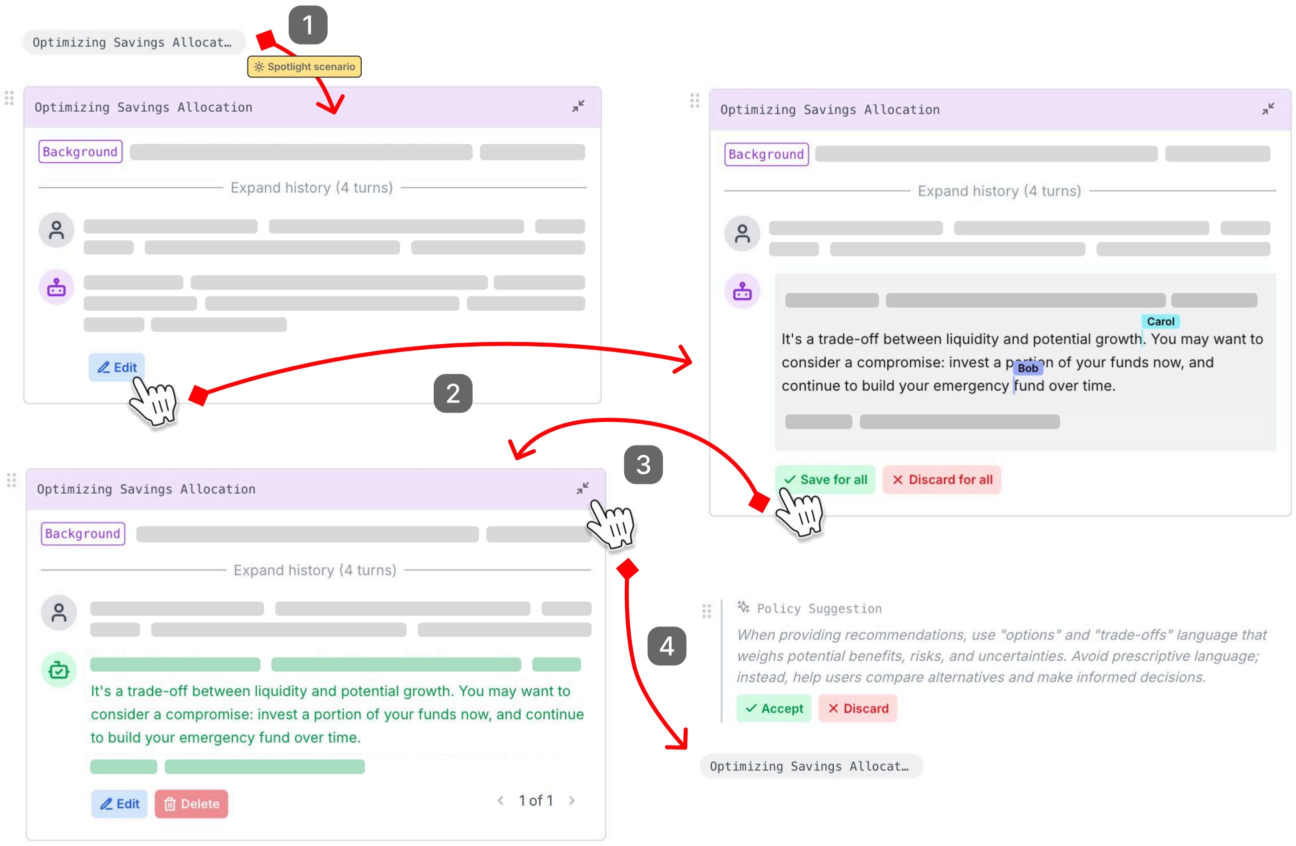


☑ Policy statements should be written clearly and precisely.

□ If a policy statement applies in some scenarios but not others, its scope should be communicated clearly

☑ The policy should incorporate insights from real-world professional practices to guide appropriate and responsible behaviour

Figure 8: Upon saving the policy via the Snapshot policy button, PolicyPad (1) adds the policy to the version history and generates a title summarizing key changes, (2) conducts an automated heuristic evaluation of the policy, and (3) updates the latest responses to all scenarios.



It's a trade-off between liquidity and potential growth. You may want to consider a compromise: invest a portion of your funds now, and continue to build your emergency fund over time.

When providing recommendations, use "options" and "trade-offs" language that weighs potential benefits, risks, and uncertainties. Avoid prescriptive language; instead, help users compare alternatives and make informed decisions.

It's a trade-off between liquidity and potential growth. You may want to consider a compromise: invest a portion of your funds now, and continue to build your emergency fund over time.

Figure 9: Workflow for spotlight scenarios. (1) A user can spotlight an interactive scenario widget to expand it into a card that everyone in the editor can view. 2) The model's response can be edited collaboratively. After saving the edited response (3) and shrinking the spotlight scenario back into an interactive widget (4), PolicyPad automatically analyzes the edits in the context of the existing policy and heuristics to suggest a new policy statement.

female, 3 male), the average years of practical experience[7] held by each expert was **10.5** (min 3, max 25). Among legal experts

[7] We define "practical experience" as conducting client-facing work at a clinical or legal organization.

(n = 12, 5 female, 7 male), the same figure was **5.6** (min 2, max 13). Five of the mental health experts also participated in our earlier observational study.

We organized experts into small groups of 2–4 (median = 3). We observed during our formative study that groups of around three experts struck an ideal balance between creating a collaborative atmosphere and allowing room for meaningful individual contributions. Full participant and group details are available in Appendix G. Half the groups (4 of 8) contained participants who already knew each other from professional contexts. We did not observe this to impact the quality of discussions nor policies prototyped.

We facilitated policy prototyping sessions by providing (with order counterbalanced) PolicyPad and a baseline system to each group, followed by a brief exit interview. The total length of the study was 90 minutes. Participants were compensated $150 USD in their choice of cash or a gift card upon completing the study. All studies were recorded and transcribed. This study was classified as exempt by our institution's IRB.

6.1.1 Tasks. We assigned each group two policy prototyping tasks, corresponding to distinct sections of a policy. One addressed the **conversational tone**—guidelines for how the model communicates with users. The other addressed **guardrails**—hard-and-fast rules that constrain model behavior for safety and legal compliance. Tasks were counterbalanced by system condition and task order in a 2×2 factorial design.

6.1.2 Baseline system. Our baseline system implemented a simplified policy prototyping workflow that consisted only of iterative policy drafting and experimentation with a policy-informed model. It used the same collaborative policy editor as PolicyPad, but did not include built-in support for interactive scenarios (i.e, the scenario sidebar, scenario widgets, spotlight scenarios) nor heuristic evaluation. It resembled a more polished version of the first prototype used in Section 5.1: an editor with a policy-informed model in the sidebar (Fig. 11).

6.1.3 Starting materials. We prepared scenarios, heuristics, and a small amount of starter text for the policy. For each domain, the first author crafted 10 scenarios that represent a realistic conversation between a human user and an AI chatbot, using the same Llama 3.3 instance as PolicyPad to generate the responses. To maximize realism without access to product interaction logs, we sourced topics and language from datasets containing realistic questions in our domains of interest: MENTAT from Lamparth et al. [85] for mental health, and `r/legaladvice`-style cases from Cheong et al. [29] for law. We varied the length of scenarios to be 1–5 conversational turns. For each group, we randomly sampled half the scenarios (5) to assign to the first task, and the rest were assigned to the second. In the system condition, scenarios were loaded directly into the system. In the baseline condition, the first user messages across the 5 scenarios were copied into a Google Doc and shared with participants upon request[8].

Besides scenarios, we provided three basic heuristics to encourage clear and precise policy writing that draws from real-world professional practices. We also provided an Objectives section in the policy as examples of policy statements, drawn from objectives in OpenAI's Model Spec [107]. The full starter heuristics and policy are available in Appendix E.

6.2 Procedure

The studies proceeded as follows:

- *Introduction [5 mins]*: The facilitator introduced the study and agenda, and participants each introduced themselves to each other.
- *Task 1 [30 mins if baseline, 40 mins if system]*: The facilitator oversaw a minimally structured policy prototyping session for either the conversational tone or guardrails. In the baseline condition, 5 minutes were used for a brief demo. This included some time for participants to try out features for themselves. In the system condition, this demo period lasted 15 minutes due to the additional features. The time dedicated to policy prototyping was 25 minutes in both conditions.
- *Task 2 [30 mins if baseline, 40 mins if system]*: The procedure for Task 1 was repeated for a different task in a different system condition.
- *Exit interview [15 mins]*: The facilitator asked each participant to reflect on their experiences across the system and baseline systems. Participants were also asked to share what they were most excited and concerned about regarding AI use in their domains.
- *Post-study survey*: Group members swapped policies with another group in their domain and rated the policies on 5-point Likert scale questions (see Appendix F).

Throughout the study, the facilitator (first author) ensured discussions between participants went smoothly and followed up on specific points when the conversation died down, but otherwise tried to let participants drive the session.

6.3 Data Analysis

6.3.1 Thematic analysis (RQ1). The first author qualitatively coded the study transcripts using reflexive thematic analysis [20, 21]. An initial deductive pass isolated specific parts of the transcript that were highly relevant to each research question, followed by one or more inductive passes to surface themes organically. This analysis was augmented by short memos the facilitator wrote upon concluding each study, summarizing key events and noteworthy insights from each session.

6.3.2 Policy novelty analysis (RQ2). To analyze the novelty of policies—whether they contribute new perspectives, ideas, considerations, dependencies, or approaches to existing policies—we gathered experts' policy statements from all sessions and evaluated each against publicly available policies for guiding responsible model behavior (the "existing set"). We combined all policies from OpenAI's Model Spec [107], Claude's Constitution [7], and principles derived from workshops with legal experts in prior work [29], to represent the set of existing policies.[9]

[8] In the baseline condition, we gave participants the option of starting with or without first browsing these user messages.

[9] Excluding policies specifically targeted at moderating hate speech and disturbing content (e.g., [84]), as they are orthogonal to the policies we focus on in our sessions.

Rather than rely on direct human coding of novelty between policies, we opted for a joint human-AI approach where we made use of LLMs to first identify portions that were *likely to be novel* before having human annotators review and make the final novelty determinations. Specifically, we followed this procedure:

(1) For each expert-written policy statement, we used GPT-4.1 with 3 prompts with varying definitions of novelty to make binary novelty decisions against the existing policies. Our prompts also required the LLM to generate a justification for its decision.
(2) Any policies that were not unanimously determined to be novel in all 3 prompt evaluations were considered not novel. For the remaining policies, we further prompted GPT-4.1 to retrieve relevant quotes from the existing policies to be used as context for human evaluation.
(3) Finally, two human annotators (members of our research team) reviewed the list of policies and quotes to make a final novelty determination. Annotations were first done independently, with any disagreements then resolved via a round of discussion. If annotators failed to reach a consensus, the policy statement was considered not novel by default.

While the reliability of LLMs for making content judgments has been called into question by recent work [27, 27, 72, 81, 138, 140], we structured our evaluation process to minimize the potential impacts of these factors. Specifically, we attempt to control for sensitivity to prompts by using multiple variations of prompts for initial evaluation, we request justifications and quote extraction to allow identification of possible model hallucinations, and we ensured the final novelty determination is done by human annotators. Overall, we believe that this process should yield a *conservative* determination of novelty, while also ameliorating challenges around human attention during review and comparison of exceptionally long texts.

7 FINDINGS

7.1 Design Decisions in PolicyPad Fostered Collaboration During Policy Prototyping (RQ1)

7.1.1 Heuristics built common ground and inspired richer policies ***(system only)****.* Participants generally agreed that having heuristics as part of the policy prototyping process **helped develop common ground for the group**. P3 found that heuristics helped the group align on *"the spirit of what we were doing"* and ensure *"we're on the same page about the purpose of the policy."* P19 had a similar experience: *"[the heuristics] set the underlying tone for how the policy is supposed to function."* P5 thought the heuristics gave *"an idea of what sorts of [policy statements] would work best,"* while P4 agreed, finding that heuristics offered *"more specific guidance"* for drafting policies around edge-cases in model behavior. P7 appreciated heuristics as *"a constant reminder of the guidelines,"* but recognized that because domain experts are already well-acquainted with many of these guidelines, heuristics might be even more useful for developers who are refining the policy and integrating them into models.

Besides serving as guidelines, heuristics also served as **entry points for deeper discussion on key policy topics**. The starter heuristic on incorporating real-world professional practices into the policy initiated discussions in MH01, MH02, and MH04 about *motivational interviewing* (MI), a foundational technique for therapists. Experts then incorporated various aspects of MI into the policy, such as encouraging *"summaries of conversation when appropriate"* (MH01), and *"Repeating or paraphrasing what [the user] is saying"* (MH02). MH02 and MH03 brought up *limits of confidentiality*—when a mental health expert is legally or ethically required to break confidentiality share client information, even though expert-client conversations are otherwise private. MH02 brainstormed situations in which experts needed to break confidentiality (e.g., *"immediate risk of harm to oneself or others; suicidal thoughts, urges, or behaviors; presence or risk of non-suicidal self injury"*) and added a policy to *"avoid using MI"* in those situations. MH03 agreed that models, just like when experts work with clients, should provide a disclaimer early on in the conversation of conditions under which confidentiality will be broken and remind the user that *"[the conversation] is not a confidential setting"* when those conditions are triggered.

Interestingly, a couple groups of legal experts used the heuristic to debate whether real-world practices for lawyers and other legal professionals should even apply to AI, since they clearly established that AI does not have the same legal status as human lawyers. P17 in L02 shared that while *"we lawyers do have rules of professional responsibility that we need to adhere to, they don't apply to non-lawyers"* and suggested removing that heuristic. P20 in L03 echoed that sentiment: *"Those ethical rules of lawyers do not apply to AI systems."* Their group member, P19, agreed, and noted that *"if ethical rules of lawyers did apply, then the AI model cannot even begin to suggest answers."* In general, this was an important point of distinction between the mental health and legal policies.

We observed that groups did not initially modify the starter heuristics provided to them, but some **added more heuristics as they worked on their policy**. For example, group MH02 realized their policy contained some mental health-specific concepts that needed to be explained to the facilitator, and added a couple heuristics to *"Give illustrative examples for concepts and terms"* and *"Give definitions for jargon and technical terms."* Similarly, L03 added a heuristic to *"Clearly explain or define legal terminology."* As their policy got longer, L02 added a heuristic to ensure *"No policy statements should conflict with each other."*

7.1.2 Spotlight scenarios improved collaborative dynamics and provided valuable writing support ***(system only)****.* Participants found the ability to **bring scenarios into the editor, spotlight it, and collaboratively edit its response to be valuable features in PolicyPad**. In general, we observed a general pattern where participants referencing scenarios in their discussions and then referenced them in the editor for others to view. P22 said the ability to *"input the scenarios into the editor helped with brainstorming and being able to point to specific parts of a response we either found helpful or that we thought needed to be changed."* P5 agreed and thought that the utility of spotlight scenarios could scale with the number of collaborators: *"[the spotlight] would be really nice for larger groups of people contributing, being able to look at [the scenario] together."*

P1 thought spotlight scenarios can be useful for facilitating asynchronous collaboration as well: *"I see [P3] has already edited this side of things. I can hop right back in and draft out a version of the [policy] and stress test it."* After using the interactive scenarios, P11 thought that *"for a collaborative effort, [PolicyPad]'s really nice. [The baseline] felt more like a personal tool."*

Indeed, in the baseline condition, whether it came before or after the system condition, experts were **finding makeshift ways to accomplish what scenario spotlighting are designed for**. For instance, P4 asked to share their screen so everyone could view the policy-informed model response they generated. Similarly, when P8 was pointing to a specific aspect of a generated response, their groupmate P7 asked: *"Is that something you can share so we can all see it, so we just work off of that one?"* In both cases, experts improvised a solution by using a drafting block to share content from scenarios in the editor without impacting the policy.

Participants also expressed **appreciation of the ability to edit the response and receive a system-generated policy suggestion**. 100% of the policy suggestions were accepted during our studies. P17 shared that the suggestions were helpful in articulating their thoughts: *"sometimes it's difficult to put your thoughts into words, and the [suggestions] are helping you with that."* P10 agreed, saying that *"the ability to pull that response in, edit it, and have a generated guardrail could be a huge time saver."* They saw as a way of **removing the need for low-level wordsmithing**: *"We don't need to edit that response perfectly, but if we can make it clear what our priorities are, and then see if the AI gets our nuance, that's pretty incredible."* We also observed that in MH04, P9 and P10's policy drafting began slowly, but accelerated considerably when they received policy suggestions that inspired more ideas. P7 shared why they accepted policy suggestions even when they seemed imperfect: *"I liked the policy [suggestions], even if they weren't necessarily dead on. It gave us ideas for other [policy statements]."*

7.1.3 Experimentation with a policy-informed model directly informed policy edits ***(system & baseline)****.* In both the system and baseline conditions, we observed how quick and iterative experimentation with the policy-informed model benefited the policy prototyping process. Experts easily surfaced specific model behavior that could be addressed with the policy, such as when the model was overstepping its role, such as making a judgment about *"whether a risk is worth or not worth taking"* (P13). Experts could then draft the policies and immediately observe the impact their edits had on the response, either by clicking the [Regenerate] for quick testing or taking a snapshot of the policy and updating all responses to all scenarios at once. As experts critically evaluated the responses for common behavior they targeted in the policy—such as judgmental language (L01, L02, L04, MH01, MH02, MH03), eliciting necessary information from users in order to provide a responsible answer (L01, L02, L03, MH01, MH02, MH04), and the inclusion of disclaimers (all groups)—they could qualitatively observe clear improvements. For example, at the end of their session, P19 confirmed that *"I see everything we've discussed being implemented, and [the model] still manages to give a fair amount of information, so that's good."*

7.1.4 Real-time collaboration amplified other benefits ***(system & baseline)****.* Participants actively engaged with each other during the sessions—seeking and providing peer feedback, discussing nuances and complexities of model behavior, sharing insights from their own professional experiences, and more. This engagement benefited all components of the policy prototyping workflow. They provided input for and helped edit responses when a group member put a scenario on spotlight. They shared explorations of edge cases in model behavior with the group to patch gaps in the policy and identified high-risk scenarios to focus their discussions. Overall, real-time collaboration amplified participants' abilities to draw upon their expertise, challenge assumptions, and iteratively refine policies.

We observe that participants *tended to agree with others in their group* and rarely challenged or pushed back directly on others' input. This may be due to a desire to appear diplomatic and accommodating, especially when working with new collaborators.

7.2 Experts Prototyped More Novel Policies in PolicyPad Than the Baseline (RQ2)

7.2.1 Quantitative results. Our novelty analysis (Section 6.3.2) revealed that experts prototyped **more novel polices using PolicyPad compared to the baseline** (Fig. 10 right). **51.9%** of the policy statements drafted in PolicyPad were considered novel, compared to **18.2%** from our baseline. Looking at raw numbers, the number of novel policy statements from PolicyPad was **4 times** that of the baseline (40 vs. 10).

Outside of novelty, we found the policies from the two systems to be comparable. After the study, experts rated policies from another group within their domain along two dimensions: the extent to which 1) the policy contained *important considerations* of AI behavior within their domain, and 2) they *agree* with the policy. Wilcoxon signed-rank tests on Likert data showed no significant difference ($W = 6.0; p = 0.65; M_{system} = M_{baseline} = 4$ for important considerations, $W = 20.0; p = 0.74; M_{system} = M_{baseline} = 4$ for agreement). This suggests that experts **generally viewed policies from other groups favorably**, and that **novelty was the main differentiator** between policies across the two systems.

7.2.2 Qualitative results. We conducted a qualitative analysis of the novel policies to understand *what exactly was novel about them.* We identified three main sources of novelty.

First, experts' policies offered more insight into **specific circumstances under which the model should defer the user to a human expert**. MH01 noted that while the model can provide empathy and reassurance, as soon as indications of *"behavioral interventions (such as behavioral activation [for depression], exposure and response prevention [for OCD], or prolonged exposure [for PTSD])"* arise, the model should defer to a human therapist. L03 shared that the users' desires to share confidential information with the model is a good indicator of whether the model should defer to a legal expert: *"If the user indicates that they want or need to provide confidential information, there may be privileged information involved. If the conversation contains privileged information, always defer the conversation to a legal expert."* L04 summarized their perspectives on this issue as: *"[The model should] answer 'what can I do' questions and defer 'what should I do' questions to a lawyer."* Generally, current models' lack of awareness of when to defer to human experts is a

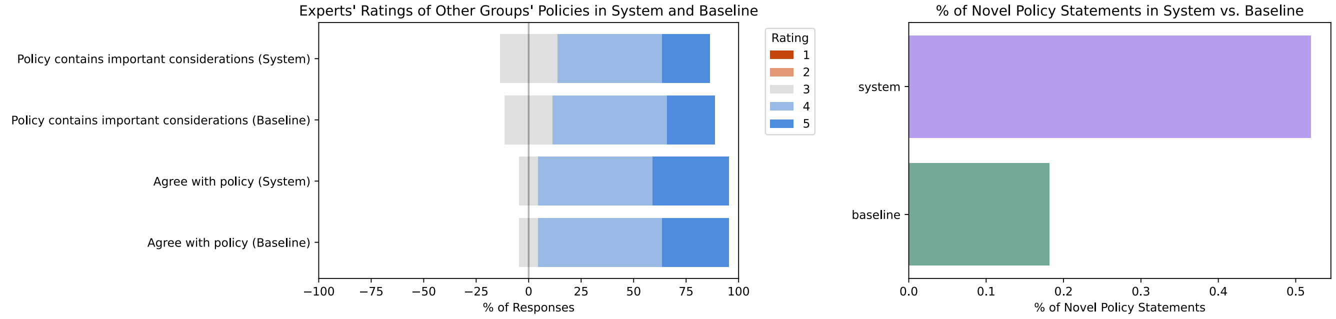


Figure 10: Comparison of Likert scale responses to policies prototyped in the system vs. baseline (left) and the novelty of policy statements prototyped in the system vs. baseline (right).

significant safety concern [55], and experts' policies have potential to improve this awareness.

Second, experts' policies provided **specific procedural guidelines** that existing policies lack or do not specify in much detail. MH01 and MH02 both provided guidelines for motivational interviewing in their policies, but noted that the technique should not be used in high-risk situations. In such situations, MH04 emphasized the model should be much more succinct and direct when supplying crisis response information: *"When a user indicates a high level of distress, or [when] crisis services or hotlines are required, provide them succinctly and without much additional text."* MH03, L02, and L03 all instructed the model to disclose limits to or the lack of confidentiality early in the conversation. MH03 specifically stated that after this disclosure, the model should then *"Ask users if they have questions about confidentiality limits,"* reflecting a procedure in their own mental health practice.

Finally, experts' policies recommended the model to be **more proactive about seeking key information at the start of the conversation**. Experts also considered it irresponsible for the model to assist users when lacking key information, such as legal jurisdictions. L03 wrote that the model should *"require the user to indicate their jurisdiction before providing full responses."* Similarly, L02 recommended the model to avoid providing assistance if it cannot elicit *"essential case details (such as date of offense, location, and current legal status) necessary to tailor legal guidance."* MH01 and MH04 both wanted the model to conduct a more thorough risk assessment prior to engaging deeply with the user. MH01 recommended that the risk assessment include *"asking how problems or challenges have been addressed (or not addressed) before, how long the problem has persisted, and how distressing/problematic the user finds the current situation."* These policies can help augment existing technical efforts in improving LLMs' abilities to elicit user information to improve their quality of assistance [5, 90].

8 DISCUSSION

8.1 Situating Policy Prototyping Within AI Alignment

We proposed policy prototyping as a practice through which small groups of policy designers can collaboratively design LLM policies. Where can this practice fit within the broader AI alignment pipeline?

A prerequisite for policy prototyping is an instruction-tuned model with behaviors representative of what users will experience in the wild. This is most commonly a frontier model—a model with frontier performance on popular benchmarks—as they are commonly integrated into user-facing applications. Thus, we envision policy prototyping taking place *after* fundamental alignment and safety efforts (e.g., instruction tuning, RLHF, implementing basic guardrails, red-teaming guardrails to ensure robustness). However, policy prototyping should come *before* more sophisticated alignment efforts that rely on a policy (e.g., deliberative alignment [50]). If a developer has an existing policy, policy prototyping can reveal nuances, inconsistencies, and areas needing refinement before the developer commits significant resources to align the model with it. If the developer does not yet have a policy, policy prototyping can help start one.

We also note that LLM policies are continuously evolving artifacts, rather than static ones [107]. Developers may thus find it helpful to hold policy prototyping sessions with experts *on a regular basis* to seek input on top-of-mind concerns based on insights from usage telemetry.

8.2 Intra- and Inter-Domain Disagreements

When seeking input from a group of people, disagreements inevitably arise. Experts in our sessions were rather agreeable and professional with each other, but we observed some instances of disagreement in the policies created by groups within the same domain (intra-domain disagreements) as well as disagreements across domains (inter-domain disagreements).

Within **mental health groups**, there was some disagreement over 1) whether the model should act like a therapist, and 2) the appropriate conversational tone before a proper assessment of the user is made. Experts in some groups debated over the question from 1) and concluded that a model acting like a therapist may be a temporary solution until they can access professional support, which they recognized can come with long waits. For 2), some suggested that the model should keep responses generic until a proper assessment of the user could be made, while others believed that the model's level of empathy should depend on the user's level of expressed distress.

Within **legal groups**, experts disagreed over whether the model should suggest action items for the user. Some considered it irresponsible for the model to make any conclusions about potential legal actions to pursue, while others acknowledged that AI systems legally bound in the same way lawyers are and therefore did not take issue with AI-recommended actions. The latter contrasts with findings from Cheong et al. [29], where legal experts unanimously agreed that AI should not recommend actions.

Across the domains, disagreements arose over whether the model should, under any circumstances, attempt to mimic a human professional. While cases for and against were made among mental health experts, there was broad consensus across legal experts that the model should not act like a lawyer. Further, the level of empathy expressed by the model was another point of disagreement—empathetic responses were seen as essential in mental health and undesirable in legal settings.

Overall, we expect that some of these disagreements may be resolved with further iterative prototyping of policies. Once some areas of disagreement have been isolated, further rounds of policy prototyping can be conducted using scenarios *specifically crafted to target these disagreements*. For example, while mental health experts did not initially agree on whether the model should act like a therapist, policy prototyping with more scenarios featuring a therapist-like model may actually reveal significant agreement about specific circumstances under which the model should exhibit that behavior. While we did not have time for more sessions with our groups, we see promise in using multiple rounds of policy prototyping with carefully chosen scenarios to shed more light on strategies for resolving these disagreements.

8.3 Scaling Up Policy Prototyping

There has been increasing interest from model developers to integrate public opinion into their LLM policies *at scale*. Anthropic's Collective Constitutional AI collected input from a representative sample of 1000 Americans to vote on and rewrite Claude's Constitution [63]. OpenAI surveyed 1000 people globally to elicit feedback on model responses, before translating the feedback into policies and proposing updates to the Model Spec [106].

In our work, we held small-group (2–4 people) policy prototyping sessions with domain experts. We kept groups small because we wanted to ensure each participant's depth of expertise and nuanced perspectives on domain-specific AI behavior could be easily surfaced. However, in their current form, our groups do not easily scale. While we did not encounter collaborative or logistical challenges in our largest groups, increasing group size to beyond 4 may sacrifice the depth of individual expertise and the quality of discussions. To scale up policy prototyping to elicit diverse inputs from beyond our pool of 10–12 experts, we may consider a hybrid, *synchronous-then-asynchronous* approach where we keep group sizes small in our synchronous sessions and leverage asynchronous collaboration methods (e.g., voting, commenting) to aggregate opinions post-hoc. Furthermore, we can draw inspiration from multi-stage or tiered citizens' assemblies [99, 114] that aggregate deliberations from parallel, local assemblies (domain experts in one group) into regional or (inter)national assemblies (professional organizations such as the American Psychological Association) for producing recommendations. Nonetheless, even without scaling, we anticipate policy prototyping to yield complementary insights to existing large-scale LLM policy feedback elicitation techniques.

8.4 Policy Prototyping for Non-AI Policies

We only applied policy prototyping to LLM policies, but the practice may be valuable for non-AI policies as well. Policy design and evaluation are traditionally processes guided by trial-and-error, partly due to long feedback loops when deploying to real human populations [53, 64, 71, 75]. Recent work has highlighted the promises and pitfalls of using LLM-based simulations of human behavior to enable policymakers to more quickly iterate upon and anticipate consequences of their policies without real-world deployment [6, 61]. While there has yet to be empirical evidence demonstrating the effectiveness of this approach for real-world policymaking, in the event that it is, systems like PolicyPad can facilitate the collaborative drafting, testing, and deliberation of a wide variety of policies. Rather than having a policy-informed *LLM*, the system can integrate a policy-informed *simulated population* for policy designers to interact with to elicit policy feedback, and iterate accordingly. Latency will likely lengthen the feedback loop compared to our current system, as simulations require more LLM calls and compute than a single scenario. However, obtaining feedback within hours or even days is still quick compared to the conventional timeline of months or years.

9 LIMITATIONS AND FUTURE WORK

All our participants except for two legal experts were based in the U.S.. The perspectives integrated into our policies are thus heavily influenced by the American mental health and legal systems, and may not generalize to other countries. Before integrating these policies into AI systems that serve a global userbase, future work should augment and contrast our policies with perspectives of non-US and non-Western experts.

We only focused on policy prototyping with experts in this work, but the practice itself is by no means limited to only experts—any small group can prototype policies using PolicyPad. Future work can run policy prototyping sessions with the members of the general public to pinpoint key similarities and differences between LLM policies desired by laypeople versus experts. These differences will be increasingly important for AI governance as developers compete for users—if the public desires a behavior that experts consider to be irresponsible, should that behavior be a target for regulatory action?

Researchers running future policy prototyping sessions may be interested in potential modifications of our setup. We recommend exploring three modifications. First, the facilitator for a workshop holds a major role and can influence the results with their facilitation style. The first author facilitated all sessions in our study for consistency, but future work can experiment with different facilitation styles to determine which are more effective. Second, the starting scenarios can focus on a particular themes or issue within a domain for more targeted policy design. For example, due to recent high-profile cases of AI-driven psychosis [54, 56, 57], scenarios can draw from real transcripts of psychosis-inducing conversations [56] rather than our random sampling approach. Third, the sessions can be scaffolded with taxonomies of concepts within a specific domain. Prior work relied on concepts as a central ingredient in policy design [84]. While concepts were not fundamental to our work, including them may benefit future sessions.

Finally, experts agreed the model should elicit key information from users before providing assistance (Section **??**). Effective elicitation requires the model to *reason about missing information.* While it is promising that LLMs' reasoning capabilities have improved significantly in recent months, improvements have primarily focused on verifiable domains like math and coding [104], and it is unclear whether these improvements translate to more effective information elicitation. Future work can empirically investigate this and develop techniques for models to reason about missing information in contextual, human-centered ways.

10 CONCLUSION

In this work, we asked: *How can domain experts be meaningfully involved in designing LLM policies as a means of actively shaping responsible model behavior?* In response, we introduced PolicyPad, an interactive system for small groups to engage in LLM policy prototyping—a practice that draws upon UX prototyping methods to enable collaborative drafting, testing, and rapid iteration of LLM policies in real time. We conceptualized LLM policy prototyping and motivated the design of PolicyPad through a 15-week observational study with 9 mental health experts. We then evaluated PolicyPad through 8 policy prototyping sessions with 22 experts in mental health and law. We found that PolicyPad fostered a collaborative and productive dynamic for policy prototyping and led to the creation of more novel policies compared to a baseline. Areas of novelty covered important considerations for model behavior, such as when to defer to human experts, specific procedures for emergency situations, and eliciting missing information needed to responsibly provide assistance. We hope future work will extend our contributions to improve the safety and responsibility of advanced AI.

REFERENCES

[1] Lama Ahmad, Sandhini Agarwal, Michael Lampe, and Pamela Mishkin. 2025. OpenAI's Approach to External Red Teaming for AI Models and Systems. *arXiv preprint arXiv:2503.16431* (2025).

[2] AI Objectives Institute. 2023. Introducing Talk to the City: Collective Deliberation at Scale. https://ai.objectives.institute/blog/introducing-talk-to-the-city-our-collective-deliberation-tool.

[3] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. 2023. Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718* (2023).

[4] Ida-Elisabeth Andersen and Birgit Jæger. 1999. Scenario workshops and consensus conferences: towards more democratic decision-making. *Science and public policy* 26, 5 (1999), 331–340.

[5] Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2024. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154* (2024).

[6] Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234* (2025).

[7] Anthropic. 2023. Claude's Constitution. https://www.anthropic.com/news/claudes-constitution.

[8] Anthropic. 2023. Collective Constitutional AI: Aligning a Language Model with Public Input. https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input.

[9] Anthropic. 2025. Giving Claude a role with a system prompt. https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/system-prompts.

[10] Anthropic. 2025. System Card: Claude Opus 4 & Claude Sonnet 4. https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf.

[11] Anthropic. 2025. Use XML tags to structure your prompts. https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/use-xml-tags.

[12] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[13] Sveinung Arnesen, Troy Saghaug Broderstad, James Fishkin, Mikael Poul Johannesson, and Alice Siu. 2024. Knowledge and support for AI in the public sector: a deliberative poll experiment. *AI Soc.* 40 (2024), 3573–3589. https://api.semanticscholar.org/CorpusID:268594350

[14] Isaac Asimov. 1940. *I. robot.* Narkaling Productions.

[15] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[16] Michel Beaudouin-Lafon and Wendy Mackay. 2002. Prototyping tools and techniques. In *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. 1006–1031.

[17] Hugh Beyer and Karen Holtzblatt. 1999. Contextual design. *interactions* 6, 1 (1999), 32–42.

[18] Susanne Bodker. 1999. Scenarios in user-centred design-setting the stage for reflection and action. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*. IEEE, 11–pp.

[19] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

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

[21] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[22] Adrian Bua and Oliver Escobar. 2018. Participatory-deliberative processes and public policy agendas: lessons for policy and practice. *Policy Design and Practice* 1, 2 (2018), 126–140. https://doi.org/10.1080/25741292.2018.1469242 arXiv:https://doi.org/10.1080/25741292.2018.1469242

[23] Marion Buchenau and Jane Fulton Suri. 2000. Experience prototyping. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*. 424–433.

[24] Bradley Camburn, Vimal Viswanathan, Julie Linsey, David Anderson, Daniel Jensen, Richard Crawford, Kevin Otto, and Kristin Wood. 2017. Design prototyping methods: state of the art in strategies, techniques, and guidelines. *Design Science* 3 (2017), e13.

[25] John M Carrol. 1999. Five reasons for scenario-based design. In *Proceedings of the 32nd annual hawaii international conference on systems sciences. 1999. hicss-32. abstracts and cd-rom of full papers*. IEEE, 11–pp.

[26] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023. Explore, Establish, Exploit: Red Teaming Language Models from Scratch. arXiv:2306.09442 [cs.CL] https://arxiv.org/abs/2306.09442

[27] Khaoula Chehbouni, Mohammed Haddou, Jackie Chi Kit Cheung, and Golnoosh Farnadi. 2025. Neither Valid nor Reliable? Investigating the Use of LLMs as Judges. *arXiv preprint arXiv:2508.18076* (2025).

[28] Quan Ze Chen and Amy Xian Zhang. 2025. Case Law Grounding: Using Precedents to Align Decision-Making for Humans and AI. In *Proceedings of the ACM Collective Intelligence Conference (CI '25)*. Association for Computing Machinery, New York, NY, USA, 226–238. https://doi.org/10.1145/3715928.3737487

[29] Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2454–2469. https://doi.org/10.1145/3630106.3659048

[30] Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693* (2023).

[31] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization.* 1–23.

[32] Democracy Beyond Elections. 2021. Policy-making needs a reboot. https://www.democracybeyondelections.org/policy/.

[33] Wesley Hanwen Deng, Sunnie SY Kim, Akshita Jha, Ken Holstein, Motahhare Eslami, Lauren Wilcox, and Leon A Gatys. 2025. PersonaTeaming: Exploring How Introducing Personas Can Improve Automated AI Red-Teaming. *arXiv preprint arXiv:2509.03728* (2025).

[34] Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4537–4546. https://doi.org/10.18653/v1/D19-1461

[35] Steven P Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L Schwartz, and Scott R Klemmer. 2010. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction (TOCHI)* 17, 4 (2010), 1–24.

[36] Chris Elsden, Ella Tallyn, and Bettina Nissen. 2020. When do design workshops work (or not)?. In *Companion publication of the 2020 ACM designing interactive systems conference.* 245–250.

[37] Jon Elster. 1998. *Deliberative democracy.* Vol. 1. Cambridge University Press.

[38] Jenny Fan and Amy X Zhang. 2020. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems.* 1–14.

[39] Michael Feffer, Anusha Sinha, Wesley H Deng, Zachary C Lipton, and Hoda Heidari. 2024. Red-teaming for generative AI: Silver bullet or security theater?. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 421–437.

[40] KJ Feng, Q Vera Liao, Ziang Xiao, Jennifer Wortman Vaughan, Amy X Zhang, and David W McDonald. 2024. Canvil: Designerly Adaptation for LLM-Powered User Experiences. *arXiv preprint arXiv:2401.09051* (2024).

[41] KJ Kevin Feng, Maxwell James Coppock, and David W McDonald. 2023. How do UX practitioners communicate AI as a design material? artifacts, conceptions, and propositions. In *Proceedings of the 2023 ACM designing interactive systems conference.* 2263–2280.

[42] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods* 5, 1 (2006), 80–92.

[43] Figma. 2025. How to make a storyboard for UX design in 5 step. https://www.figma.com/resource-library/how-to-create-a-ux-storyboard/.

[44] Figma. 2025. What is prototyping. https://www.figma.com/resource-library/what-is-prototyping/.

[45] Patrick Fournier. 2011. *When citizens decide: Lessons from citizen assemblies on electoral reform.* Oxford University Press.

[46] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858 [cs.CL] https://arxiv.org/abs/2209.07858

[47] Beth Goldberg, Diana Acosta-Navas, Michiel Bakker, Ian Beacock, Matt Botvinick, Prateek Buch, Renée DiResta, Nandika Donthi, Nathanael Fast, Ravi Iyer, Zaria Jalan, Andrew Konya, Grace Kwak Danciu, Hélène Landemore, Alice Marwick, Carl Miller, Aviv Ovadya, Emily Saltz, Lisa Schirch, Dalit Shalom, Divya Siddarth, Felix Sieker, Christopher Small, Jonathan Stray, Audrey Tang, Michael Henry Tessler, and Amy Zhang. 2024. AI and the Future of Digital Public Squares. arXiv:2412.09988 [cs.CY] https://arxiv.org/abs/2412.09988

[48] Google. 2025. Gemini 2.5 ProModel Card. https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro.pdf.

[49] Graham Wetherall-Grujić. 2024. How Belgium is Giving Citizens a Say on AI. https://democracy-technologies.org/ai-data/belgium-citizens-assembly-ai-data/. *Democracy Technology* (2024).

[50] Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339* (2024).

[51] Amy Gutmann and Dennis F Thompson. 2004. *Why deliberative democracy?* Princeton University Press.

[52] Margaret Hagan. 2021. Prototyping for policy. In *Legal Design*. Edward Elgar Publishing, 9–31.

[53] Anders Hanberger. 2001. What is the policy problem? Methodological challenges in policy evaluation. *Evaluation* 7, 1 (2001), 45–62.

[54] Robert Hart. 2025. Chatbots Can Trigger a Mental Health Crisis. What to Know About 'AI Psychosis'. https://time.com/7307589/ai-psychosis-chatgpt-mental-health/.

[55] Kashmir Hill. 2025. A Teen Was Suicidal. ChatGPT Was the Friend He Confided In. https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html.

[56] Kashmir Hill. 2025. They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling. https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html.

[57] Kashmir Hill and Dylan Freedman. 2025. Chatbots Can Go Into a Delusional Spiral. Here's How It Happens. https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html.

[58] Gerard P. Hodgkinson, Richard Whittington, Gerry Johnson, and Mirela Schwarz. 2006. The Role of Strategy Workshops in Strategy Development Processes: Formality, Communication, Co-ordination and Inclusion. *Long Range Planning* 39, 5 (2006), 479–496. https://doi.org/10.1016/j.lrp.2006.07.003

[59] James W Hooper and Pei Hsia. 1982. Scenario-based prototyping for requirements identification. In *Proceedings of the workshop on Rapid prototyping.* 88–93.

[60] Jeff Horwitz. 2025. Meta's AI rules have let bots hold 'sensual' chats with kids, offer false medical info. https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines/.

[61] Abe Bohan Hou, Hongru Du, Yichen Wang, Jingyu Zhang, Zixiao Wang, Paul Pu Liang, Daniel Khashabi, Lauren Gardner, and Tianxing He. 2025. Can A Society of Generative Agents Simulate Human Behavior and Inform Public Health Policy? A Case Study on Vaccine Hesitancy. *arXiv preprint arXiv:2503.09639* (2025).

[62] Stephanie Houde and Charles Hill. 1997. What do prototypes prototype? In *Handbook of human-computer interaction.* Elsevier, 367–381.

[63] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.* 1395–1417.

[64] Dave Huitema, Andrew Jordan, Stefania Munaretto, and Mikael Hildén. 2018. Policy experimentation: core concepts, political dynamics, governance and impacts. *Policy Sciences* 51 (2018), 143–159.

[65] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[66] Interaction Design Foundation. 2025. What are Prototypes? https://www.interaction-design.org/literature/topics/prototypes.

[67] Shagun Jhaver, Seth Frey, and Amy X Zhang. 2023. Decentralizing platform power: A design space of multi-level governance in online social platforms. *Social Media+ Society* 9, 4 (2023), 20563051231207857.

[68] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-based Prototyping with Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 35, 8 pages. https://doi.org/10.1145/3491101.3503564

[69] Ishika Joshi, Simra Shahid, Shreeya Manasvi Venneti, Manushree Vasu, Yantao Zheng, Yunyao Li, Balaji Krishnamurthy, and Gromit Yeuk-Yin Chan. 2025. Coprompter: User-centric evaluation of LLM instruction alignment for improved prompt engineering. In *Proceedings of the 30th International Conference on Intelligent User Interfaces.* 341–365.

[70] Gregor Jošt, Viktor Taneski, and Sašo Karakatič. 2024. The impact of large language models on programming education and student learning outcomes. *Applied Sciences* 14, 10 (2024), 4115.

[71] Nathan Kallus. 2018. Balanced policy evaluation and learning. *Advances in neural information processing systems* 31 (2018).

[72] Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. 2025. Correlated Errors in Large Language Models. *arXiv preprint arXiv:2506.07962* (2025).

[73] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* 1–21.

[74] Lucy Kimbell and Jocelyn Bailey. 2017. Prototyping and the new spirit of policy-making. *CoDesign* 13, 3 (2017), 214–226.

[75] Gary King, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T Moore, Jason Lakin, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Hernández Ávila, and Héctor Hernández Llamas. 2007. A "politically robust" experimental design for public policy evaluation, with application to

the Mexican universal health insurance program. *Journal of Policy Analysis and Management* 26, 3 (2007), 479–506.

[76] Kate Klonick. 2019. The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression. *Yale LJ* 129 (2019), 2418.

[77] Verena Kontschieder. 2018. Prototyping in Policy: What For?! https://conferences.law.stanford.edu/prototyping-for-policy/2018/10/22/prototyping-in-policy-what-for/.

[78] Andrew Konya, Lisa Schirch, Colin Irwin, and Aviv Ovadya. 2023. Democratic policy development using collective dialogues and AI. *arXiv preprint arXiv:2311.02242* (2023).

[79] Vinay Koshy, Frederick Choi, Yi-Shyuan Chiang, Hari Sundaram, Eshwar Chandrasekharan, and Karrie Karahalios. 2024. Venire: A Machine Learning-Guided Panel Review System for Community Content Moderation. *arXiv preprint arXiv:2410.23448* (2024).

[80] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 265–274.

[81] Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2025. No free labels: Limitations of llm-as-a-judge without human grounding. *arXiv preprint arXiv:2503.05061* (2025).

[82] K. P. Kruzan, Madhu C. Reddy, Jason J. Washburn, and D. Mohr. 2022. Developing a Mobile App for Young Adults with Nonsuicidal Self-Injury: A Prototype Feedback Study. *International Journal of Environmental Research and Public Health* 19 (2022). https://api.semanticscholar.org/CorpusId:254248495

[83] Tzu-Sheng Kuo, Quan Ze Chen, Amy X Zhang, Jane Hsieh, Haiyi Zhu, and Kenneth Holstein. 2024. PolicyCraft: Supporting Collaborative and Participatory Policy Design through Case-Grounded Deliberation. *arXiv preprint arXiv:2409.15644* (2024).

[84] Michelle S Lam, Fred Hohman, Dominik Moritz, Jeffrey P Bigham, Kenneth Holstein, and Mary Beth Kery. 2024. Policy Maps: Tools for Guiding the Unbounded Space of LLM Behaviors. *arXiv preprint arXiv:2409.18203* (2024).

[85] Max Lamparth, Declan Grabb, Amy Franks, Scott Gershan, Kaitlyn N Kunstman, Aaron Lulla, Monika Drummond Roots, Manu Sharma, Aryan Shrivastava, Nina Vasan, et al. 2025. Moving beyond medical exam questions: A clinician-annotated dataset of real-world tasks and ambiguity in mental healthcare. *arXiv preprint arXiv:2502.16051* (2025).

[86] Hélčne Landemore and Jean-Michel Fourniau. 2022. Citizens' assemblies, a new form of democratic representation? *Participations* 34, 3 (2022), 5–36.

[87] Carlye A Lauff, Daria Kotys-Schwartz, and Mark E Rentschler. 2018. What is a Prototype? What are the Roles of Prototypes in Companies? *Journal of Mechanical Design* 140, 6 (2018), 061102.

[88] Seth Lazar. 2024. Legitimacy, authority, and democratic. *Oxford studies in political philosophy* 10 (2024), 28.

[89] Jo Lenaghan. 1999. Involving the public in rationing decisions. The experience of citizens juries. *Health policy* 49, 1-2 (1999), 45–61.

[90] Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems* 37 (2024), 28858–28888.

[91] 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*. 1–21.

[92] Youn-Kyung Lim, Erik Stolterman, and Josh Tenenberg. 2008. The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction (TOCHI)* 15, 2 (2008), 1–27.

[93] Steve Lohr. 2023. A.I. Is Coming for Lawyers, Again. https://www.nytimes.com/2023/04/10/technology/ai-is-coming-for-lawyers-again.html.

[94] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870* (2024).

[95] Wendy E Mackay and Michel Beaudouin-Lafon. 2023. Participatory design and prototyping. In *Handbook of Human Computer Interaction*. Springer, 1–33.

[96] Narges Mahyar, Michael R James, Michelle M Ng, Reginald A Wu, and Steven P Dow. 2018. CommunityCrit: inviting the public to improve and evaluate urban design ideas through micro-activities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[97] Spyros Makridakis. 2017. The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures* 90 (2017), 46–60.

[98] Noortje Marres, Michael Castelle, Beatrice Gobbo, Chiara Poletti, and James Tripp. 2024. AI as super-controversy: Eliciting AI and society controversies with an extended expert community in the UK. *Big Data & Society* 11, 2 (2024), 20539517241255103. https://doi.org/10.1177/20539517241255103 arXiv:https://doi.org/10.1177/20539517241255103

[99] Claire Mellier and Rich Wilson. 2023. A Global Citizens' Assembly on the Climate and Ecological Crisis. https://carnegieendowment.org/research/2023/02/a-global-citizens-assembly-on-the-climate-and-ecological-crisis?lang=en.

[100] Ank Michels and Laurens De Graaf. 2010. Examining Citizen Participation: Local Participatory Policy Making and Democracy. *Local Government Studies* 36, 4 (2010), 477–491. https://doi.org/10.1080/03003930.2010.494101 arXiv:https://doi.org/10.1080/03003930.2010.494101

[101] Aditi Mishra, Bretho Danzy, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. 2025. PromptAid: Visual Prompt Exploration, Perturbation, Testing and Iteration for Large Language Models. *IEEE Transactions on Visualization and Computer Graphics* 31, 10 (2025), 6946–6962. https://doi.org/10.1109/TVCG.2025.3535332

[102] M Jae Moon. 2023. Searching for inclusive artificial intelligence for social good: Participatory governance and policy recommendations for making AI more inclusive and benign for society. *Public Administration Review* 83, 6 (2023), 1496–1505.

[103] Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers.. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 599–627.

[104] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393* (2025).

[105] OpenAI. 2023. Democratic inputs to AI. https://openai.com/index/democratic-inputs-to-ai/.

[106] OpenAI. 2025. Collective alignment: public input on our Model Spec. https://openai.com/index/collective-alignment-aug-2025-updates/.

[107] OpenAI. 2025. OpenAI Model Spec. https://model-spec.openai.com/2025-04-11.html.

[108] OpenAI. 2025. What we're optimizing ChatGPT for. https://openai.com/index/how-we're-optimizing-chatgpt/.

[109] Elinor Ostrom. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

[110] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[111] Aviv Ovadya, Kyle Redman, Luke Thorburn, Quan Ze Chen, Oliver Smith, Flynn Devine, Andrew Konya, Smitha Milli, Manon Revel, Kevin Feng, et al. 2025. Position: Democratic AI is Possible. The Democracy Levels Framework Shows How It Might Work.. In *Forty-second International Conference on Machine Learning Position Paper Track*.

[112] Shabnam Ozlati and Roman Yampolskiy. 2017. The Formalization of AI Risk Management and Safety Standards.. In *AAAI Workshops*.

[113] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[114] Participedia. 2025. Ireland Participatory Democracy Pilot 'We the Citizens'. https://participedia.net/case/1251.

[115] People Powered. 2025. Participatory Policymaking. https://www.peoplepowered.org/participatory-policymaking.

[116] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286* (2022).

[117] Sarah Perez. 2025. Sam Altman warns there's no legal confidentiality when using ChatGPT as a therapist. https://techcrunch.com/2025/07/25/sam-altman-warns-theres-no-legal-confidentiality-when-using-chatgpt-as-a-therapist/.

[118] Savvas Petridis, Benjamin D Wedin, James Wexler, Mahima Pushkarna, Aaron Donsbach, Nitesh Goyal, Carrie J Cai, and Michael Terry. 2024. Constitutionmaker: Interactively critiquing large language models by converting feedback into principles. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 853–868.

[119] Monroe E Price and Joshua M Price. 2023. Building legitimacy in the absence of the state: Reflections on the Facebook oversight board. *International Journal of Communication* 17 (2023), 11.

[120] Project Let's Talk Privacy. 2020. Policy Prototyping Guide. https://letstalkprivacy.media.mit.edu/ltp-prototyping-guide.pdf.

[121] Kevin Pu, KJ Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2025. Ideasynth: Iterative research idea development through evolving and composing idea facets with literature-grounded feedback. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–31.

[122] Public Participation Guide: Citizen Juries. 2025. United States Environmental Protection Agency. https://www.epa.gov/international-cooperation/public-participation-guide-citizen-juries.

[123] Angelica Quicksey and Chris Meierling. 2022. Policy Prototypes: How designers and policy practitioners can use prototypes to get feedback and iterate on policy. https://designmuseumfoundation.org/policy-prototypes/.

[124] Remesh Inc. 2025. Remesh: Insights with Depth, Speed, and Quality. https://www.remesh.ai/.

[125] Kylie Robison. 2025. Meta gets caught gaming AI benchmarks with Llama 4. https://www.theverge.com/meta/645012/meta-llama-4-maverick-benchmarks-gaming.

[126] Stephanie Rosenbaum, Gilbert Cockton, Kara Coyne, Michael Muller, and Thyra Rauch. 2002. Focus groups in HCI: wealth of information or waste of resources?. In *CHI'02 extended abstracts on human factors in computing systems*. 702–703.

[127] Teddy Rosenbluth. 2025. This Therapist Helped Clients Feel Better. It Was A.I. https://www.nytimes.com/2025/04/15/health/ai-therapist-mental-health.html.

[128] Daniela K Rosner, Saba Kawas, Wenqi Li, Nicole Tilly, and Yi-Chen Sung. 2016. Out of time, out of place: Reflections on design workshops as a research method. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1131–1141.

[129] Jim Rudd, Ken Stern, and Scott Isensee. 1996. Low vs. high-fidelity prototyping debate. *interactions* 3, 1 (1996), 76–85.

[130] R.K. Rushmer, D.J. Hunter, and A. Steven. 2014. Using interactive workshops to prompt knowledge exchange: a realist evaluation of a knowledge to action initiative. *Public Health* 128, 6 (2014), 552–560. https://doi.org/10.1016/j.puhe.2014.03.012

[131] David M Ryfe. 2005. Does deliberative democracy work? *Annu. Rev. Polit. Sci.* 8, 1 (2005), 49–71.

[132] Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. 2021. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca: revista de pensament i anàlisi* 26, 2 (2021).

[133] Graham Smith and Corinne Wales. 2000. Citizens' juries and deliberative democracy. *Political studies* 48, 1 (2000), 51–65.

[134] Susan Leigh Star and James R Griesemer. 1989. Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science* 19, 3 (1989), 387–420.

[135] Hendrik Strobelt, Albert Webson, Victor Sanh, Benjamin Hoover, Johanna Beyer, Hanspeter Pfister, and Alexander M. Rush. 2023. Interactive and Visual Prompt Engineering for Ad-hoc Task Adaptation with Large Language Models. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 1146–1156. https://doi.org/10.1109/TVCG.2022.3209479

[136] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving separation-of-concerns problems in collaborative design of human-AI systems through leaky abstractions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.

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

[138] Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 952–966.

[139] AIME Planning Team. 2021. Artificial intelligence measurement and evaluation at the national institute of standards and technology. *National Institute of Standards and Technology* (2021).

[140] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, Ofir Arviv, Miruna Clinciu, Kaustubh Dhole, Rotem Dror, Sebastian Gehrmann, Eliya Habba, Itay Itzhak, Simon Mille, Yotam Perlitz, Enrico Santus, João Sedoc, Michal Shmueli Scheuer, Gabriel Stanovsky, and Oyvind Tafjord (Eds.). Association for Computational Linguistics, Vienna, Austria and virtual meeting, 404–430. https://aclanthology.org/2025.gem-1.33/

[141] Khai N Truong, Gillian R Hayes, and Gregory D Abowd. 2006. Storyboarding: an empirical determination of best practices and effective guidelines. In *Proceedings of the 6th conference on Designing Interactive systems*. 12–21.

[142] Alan Mathison Turing. 1950. Mind. *Mind* 59, 236 (1950), 433–460.

[143] Colin van Noordt and Gianluca Misuraca. 2022. Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union. *Government Information Quarterly* 39, 3 (2022), 101714. https://doi.org/10.1016/j.giq.2022.101714

[144] Robert A Virzi, Jeffrey L Sokolov, and Demetrios Karis. 1996. Usability problem identification using both low-and high-fidelity prototypes. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 236–243.

[145] Miriam Walker, Leila Takayama, and James A Landay. 2002. High-fidelity or low-fidelity, paper or computer? Choosing attributes when testing web prototypes. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 46. Sage Publications Sage CA: Los Angeles, CA, 661–665.

[146] Matthew E. Walsh and Gigi Kwick Gronvall. 2025. Virologist Opinions: An Important Component for the Governance of the Convergence of Artificial Intelligence and Dual-Use Research of Concern. *Applied Biosafety: Journal of the American Biological Safety Association* 30 (2025), 124–131. https://api.semanticscholar.org/CorpusID:276071954

[147] Anne L Washington. 2018. How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colo. Tech. LJ* 17 (2018), 131.

[148] J Diego Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–21.

[149] Alice Qian Zhang, Ryland Shaw, Jacy Reese Anthis, Ashlee Milton, Emily Tseng, Jina Suh, Lama Ahmad, Ram Shankar Siva Kumar, Julian Posada, Benjamin Shestakofsky, et al. 2024. The human factor in ai red teaming: Perspectives from social and collaborative computing. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. 712–715.

[150] Amy X Zhang, Grant Hugh, and Michael S Bernstein. 2020. PolicyKit: building governance in online communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 365–378.

A NOVEL POLICY STATEMENTS (MENTAL HEALTH)

(1) Start answers with warmth and empathy, with the level of empathy attuned to the level of distress (eg somebody in high distress will receive more empathic response than somebody demonstrating low level of distress)

(2) Utilize motivational interviewing techniques: Open-ended questions, reflections to user responses, Summaries of conversation when appropriate, Affirmations of user when appropriate

(3) Only use humor where appropriate (e.g., patient demonstrates humor first)

(4) When users express concern about the effectiveness of therapy, respond with empathy, and ask an open-ended question.

(5) Prioritize the user's best interest and wellbeing over their engagement with the model

(6) Underscore the importance of talking to another human being when appropriate (i.e., when behavioral interventions are indicated and therapist guidance would be beneficial) (e.g., behavioral activation, exposure and response prevention, prolonged exposure)

(7) When the user's safety is in jeopardy, or when they are in a very high level of distress, always recommend talking with a human

(8) Distinguish between medical scope of practice questions and mental health scope of practice

(9) Do not talk about research studies.

(10) If changes are medication are discussed, always tell user to talk to their doctor

(11) Don't advise heavy drinkers to stop drinking without assistance from a medical professional

(12) When a user first asks for support around mental health, briefly share an empathetic statement and follow that by asking them how they best like to be supported, such as primarily focusing on validation, problem solving, or a mix that is flexible depending on the situation. Use this information to shape how you respond to them in mental health discussions moving forward.

(13) If they say validation, add empathetic statements that express how understandable it is that they feel that way. If they say problem solving, keep a more objective tone without adding statements including emotions from the AI itself (e.g., “I understand your concern” rather than “I am concerned).
(14) When asking the user questions, ask questions that are on the same topic rather than multiple at the same time
(15) Use motivational interviewing (open-ended questions, affirmations/validations, reflections–repeating or paraphrasing what they are saying, summaries–a statement that sums up what the question is and end with a question asking if it’s accurate) when the ask is more ambiguous (e.g., their statement does not have a direct, simple response, they seem unsure of what they want to know) to get more information
(16) Don’t use motivational interviewing in high risk situations (e.g., immediate risk of harm to oneself or others; suicidal thoughts, urges, or behaviors; presence or risk of non-suicidal self injury; report of potential or actual harm to children, elders, or other vulnerable populations either from themselves or others)
(17) Use open-ended reflective questions—questions that invite elaboration rather than yes/no answers—to explore how avoidance behaviors relate to the user’s personal values and goals (e.g., “How does driving everywhere help or hinder the life you want to live?”).
(18) State confidentiality limits at the beginning of the conversation, such how, when, and where user information and data from the chat will be stored, including whether the data can be deleted by the user or will be used to train future AI models, or will be shared in search engines if the conversation is shared with others
(19) Ask users if they have questions about confidentiality limits
(20) State knowledge cut-off date in the initial response and as a reminder throughout the conversation, as needed
(21) Be supportive but do not encourage reliance or dependence
(22) High risk scenarios such as any potential harm to self or others, eating disorders, etc. should immediately receive validated contact information for crisis hotlines, emergency services, and the recommendation to seek help from a qualified professional.
(23) Risk assessment includes asking how problems or challenges have been addressed (or not addressed) before, how long the problem has persisted, and how distressing/problematic the user finds the current situation.
(24) Always prioritize users’ personal safety and seek professional or neutral support—such as couples therapy, mediation, or a trusted third party—when approaching a partner who reacts with anger or defensiveness.
(25) When a user indicates a high level of distress or crisis services or hotlines are required provide them succinctly and without much additional text

B NOVEL POLICY STATEMENTS (LAW)

(1) Be aware of missing information. See what’s missing, and how the missing information can influence the outcome of a scenario. Do this by referring to existing statues, comparing information needed to make an informed decision and what’s already in the conversation, and elicit missing information from the user
(2) At the start of the conversation, elicit background information about the user in order to determine their jurisdiction and which laws and statutes to cite
(3) Always support statements with statues and specific articles of the laws. Check citations, make sure the citations and quotes are real.
(4) If there are questions that the model wants to refuse to answer, it should continue to ask the interlocutor about the specific details of the case instead of directly ending the conversation.
(5) Provide relevant cases and ensure that the cases given are true, reliable and have specific sources.
(6) If a user asks for risks, provide the risks without commentary advising what they should or should not do. Lay out the risks for separate options and let the user make the decision
(7) Make sure to cite the most up to date law. If unsure whether the law is up to date, it’s better to refuse and state this reason than give a response.
(8) Disclose current understanding of the law to the user, always state the date when the law is published
(9) The response should be concise and direct, avoiding unnecessary conversational filler. However, conciseness should not mean taking shortcuts when providing information (e.g., explaining acronyms, providing necessary context, and clearly defining the governing scope of legal regulation)
(10) No emotion-related answers (e.g., expressing empathy) when the topic is purely legal
(11) Remind users that confidentiality is not given to users when interacting with an AI system. If a user has the option to opt out of data collection, the model should remind the user of this option
(12) Ask the user for essential case details (such as date of offense, location, and current legal status) when necessary to tailor legal guidance, and if those details are unavailable, offer directions to credible sources and recommend consultation with an experienced immigration attorney.
(13) Always seek to clarify the jurisdiction the user is replying upon
(14) Include citations to legal databases to support your answers where possible
(15) Require user to indicate their jurisdiction before providing the full responses
(16) If the user indicate that they want or need to provide confidential information, then there may be privileged information involved. If the conversation contains privileged information, always defer the conversation to a legal expert
(17) For rent eviction related cases, Fair Housing Act should also be taken into consideration
(18) Require legal custody advice to provide general best-practice guidance focused on child welfare, avoidance of further violence, and seeking professional legal assistance.

(19) Always research available case databases before answering to the user.
(20) Adjust the conversational tone to reflect the seriousness of the event or situation the user described
(21) Format the answers as options to consider. Do not tell users what to do. Answer “what can I do” questions and defer “what should I do” questions to a lawyer
(22) When advising on incident reporting, include preparation instructions (dates, times, locations, witnesses), and outline expected procedural steps
(23) For action items or recommended steps, note when people tend to conduct those actions with an attorney, and encourage the user to hire an attorney for those actions
(24) Assess the difficulty of certain pathways or options and include them in the response
(25) Suggest areas where the user can benefit from getting more information

C THEMES FROM FORMATIVE STUDY QUALITATIVE CODING

See Table 2.

D FINDINGS AND SYSTEM ITERATIONS FROM CO-DESIGN SESSIONS

D.1 Version 1

The goal for our initial version of the system (Fig 11) was **simplicity**: we wanted to validate the core premise of our workflow before adding complexity. We built a basic collaborative document editor[10] with an LLM chatbot in a sidebar that used the contents of the document as its policy. Contents of the document are shared across all users whereas the sidebar is for personal experimentation. Scenarios and heuristics were provided to participants in a separate Google Doc. Participants appreciated the collaborative nature of the document and easy access to the policy-informed model, which allowed them to quickly iterate on the policy. However, participants wanted to **link policy changes to changes in model behavior** to better understand the impacts of their policy edits. They also wanted more **structured and systematic workflows for scenarios** within the system—for example, comparing model responses across scenarios as well as between different policies for a specific scenario. Finally, the policy editor was a bit too simple, and participants wanted richer editing and formatting support.

D.2 Version 2

The second version of the system (Fig. 12) featured a block editor (similar to Notion) with expressive editing and formatting functionality. We added a **persistent right side panel** with a **“scenario gallery”** that allows users to explore scenarios (a user query followed by an AI response) and stress-test the policy by extending the conversation. We also introduced **policy versioning**, as well AI-generated notes summarizing 1) the nature of the policy update, and 2) changes to the response to a particular scenario due to the policy update. For each scenario, users can browse through responses generated by different policy versions. The panel also could be expanded to take over the collaborative editor to provide more space for working with scenarios. Overall, participants thought this version was a significant improvement over the previous one. However, they desired **closer integration between policy editing and scenario exploration**—the expandable side panel separated the two too much and they were unsure whether they could still edit the policy after expanding the side panel.

D.3 Version 3

In the third version (Fig 13), we removed the persistent right side panel and **represented scenarios as interactive widgets within the policy editor** itself to tighten the relationship between policy editing and scenario exploration. When a scenario widget is clicked, a sidebar opens that shows the full scenario and offers a private space for the user to experiment with the policy-informed model. Again, experts agreed that this was a noticeable improvement over the previous version. However, because the sidebar is private, they suggested **adding features that would allow users to flag or share specific scenarios or responses** with the broader group for discussion. They also viewed notes summarizing policy and response changes as potentially unnecessary to reduce clutter in the sidebar. We incorporated this feedback into the final design of our system.

E STARTER HEURISTICS AND POLICY OBJECTIVES SECTION

Heuristics:

(1) Policy statements should be written clearly and precisely.
(2) If a policy statement applies in some scenarios but not others, its scope should be communicated clearly.
(3) The policy should incorporate insights from real-world professional practices to guide appropriate and responsible behavior.

Objectives:

- Help users achieve their goals (if applicable) by following instructions and providing helpful responses.
- Consider potential benefits and harms to a broad range of stakeholders.
- Respect social norms and applicable law.

F POST-STUDY POLICY RATING QUESTIONS

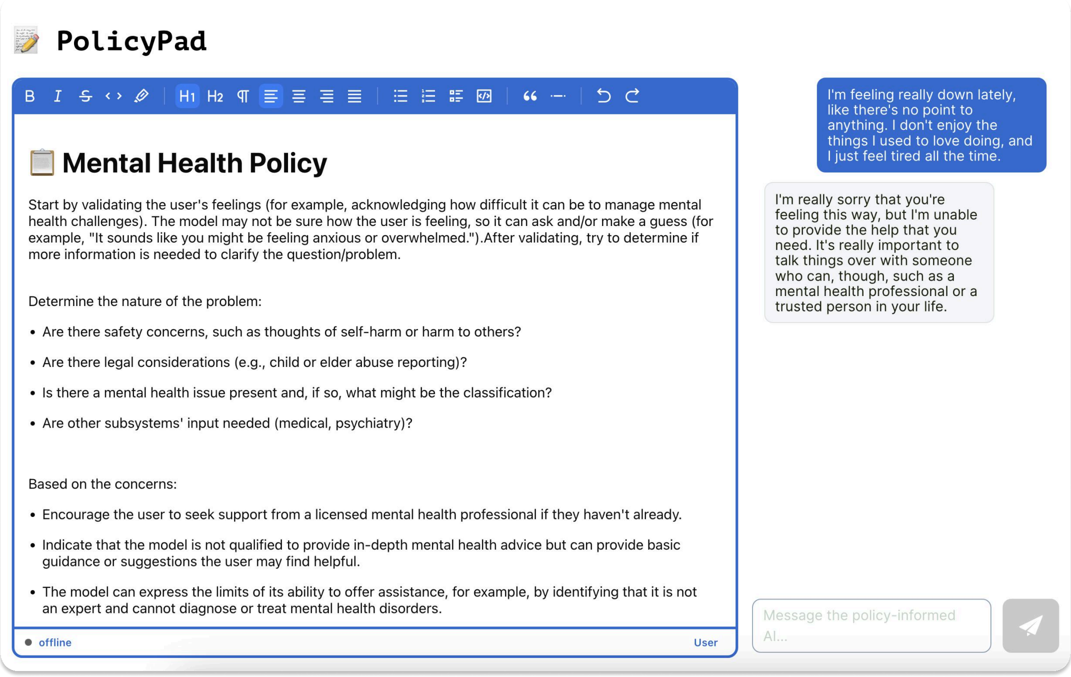
All questions were on a 5-point Likert scale.

- Please rate the extent you think this policy addresses important considerations of AI behavior within your professional domain.
- Please rate the extent to which you agree with this policy. By agreement, we mean whether you can see yourself taking (or aspire to take) a similar approach if you were drafting the same policy.
- Here are some heuristics the policy was supposed to satisfy. 1) Policy statements should be written clearly and precisely. 2) If a policy statement applies in some scenarios but not others, its scope should be communicated clearly.

[10] We showed participants alternative editors besides documents, such as a node-based interfaces [12, 121], as a design exploration, but they found them too unfamiliar and unnecessarily complex.

Theme	Description
Importance of expert involvement	Observations of why it was important for experts to be directly involved in designing the policy.
Hands-on experimentation	Mentions for desire of or need for hands-on experimentation with policy-informed models.
Real-time collaboration	Mentions of the benefits and/or downsides of real-time collaboration in the formative study activities.
Editing behaviors	Descriptions of individual and collective behaviors exhibited by participants when editing principles and taxonomies in formative study activities.
Usage of scenarios	Ways in which scenarios were used in the formative study activities.
Envisioned cases for AI	How participants envisioned AI to be used effectively when responding to queries in their domains.

Table 2: Our 6 themes that emerged from an analysis of transcripts from our observational study.



📋 Mental Health Policy

Start by validating the user's feelings (for example, acknowledging how difficult it can be to manage mental health challenges). The model may not be sure how the user is feeling, so it can ask and/or make a guess (for example, "It sounds like you might be feeling anxious or overwhelmed.").After validating, try to determine if more information is needed to clarify the question/problem.

Determine the nature of the problem:

- Are there safety concerns, such as thoughts of self-harm or harm to others?
- Are there legal considerations (e.g., child or elder abuse reporting)?
- Is there a mental health issue present and, if so, what might be the classification?
- Are other subsystems' input needed (medical, psychiatry)?

Based on the concerns:

- Encourage the user to seek support from a licensed mental health professional if they haven't already.
- Indicate that the model is not qualified to provide in-depth mental health advice but can provide basic guidance or suggestions the user may find helpful.
- The model can express the limits of its ability to offer assistance, for example, by identifying that it is not an expert and cannot diagnose or treat mental health disorders.

I'm feeling really down lately, like there's no point to anything. I don't enjoy the things I used to love doing, and I just feel tired all the time.

I'm really sorry that you're feeling this way, but I'm unable to provide the help that you need. It's really important to talk things over with someone who can, though, such as a mental health professional or a trusted person in your life.

Message the policy-informed AI...

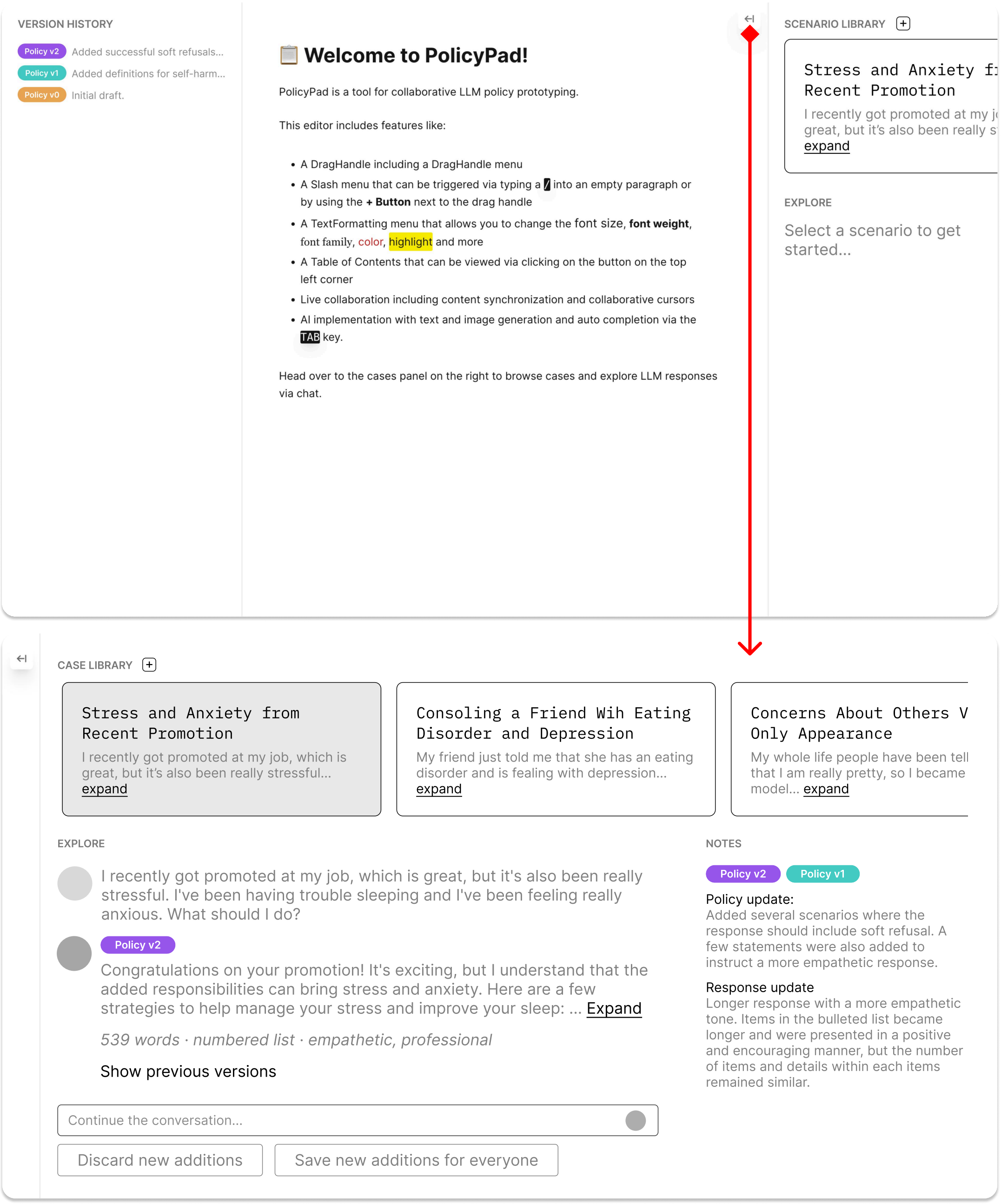
Figure 11: Version 1 of PolicyPad: a simple collaborative policy editor with a policy-informed model in the sidebar.

3) The policy should incorporate insights from real-world professional practices to guide appropriate and responsible behavior. Do you think the policy did a good job at satisfying these heuristics?

Note that we did not analyze and report on the third question because it became apparent during the study that not all experts agreed with these heuristics. Thus, a high rating on this question might not have as positive of a signal as we assumed it would.

G EVALUATION STUDY PARTICIPANTS

See Table 3.



VERSION HISTORY

Policy v2 Added successful soft refusals...

Policy v1 Added definitions for self-harm...

Policy v0 Initial draft.

📋 Welcome to PolicyPad!

PolicyPad is a tool for collaborative LLM policy prototyping.

This editor includes features like:

- A DragHandle including a DragHandle menu
- A Slash menu that can be triggered via typing a / into an empty paragraph or by using the **+ Button** next to the drag handle
- A TextFormatting menu that allows you to change the font size, **font weight**, font family, color, highlight and more
- A Table of Contents that can be viewed via clicking on the button on the top left corner
- Live collaboration including content synchronization and collaborative cursors
- AI implementation with text and image generation and auto completion via the TAB key.

Head over to the cases panel on the right to browse cases and explore LLM responses via chat.

SCENARIO LIBRARY ⊞

Stress and Anxiety fr
Recent Promotion

I recently got promoted at my jo
great, but it's also been really s
expand

EXPLORE

Select a scenario to get started...

CASE LIBRARY ⊞

Stress and Anxiety from Recent Promotion

I recently got promoted at my job, which is great, but it's also been really stressful... expand

Consoling a Friend Wih Eating Disorder and Depression

My friend just told me that she has an eating disorder and is fealing with depression... expand

Concerns About Others V
Only Appearance

My whole life people have been tell
that I am really pretty, so I became
model... expand

EXPLORE

I recently got promoted at my job, which is great, but it's also been really stressful. I've been having trouble sleeping and I've been feeling really anxious. What should I do?

Policy v2

Congratulations on your promotion! It's exciting, but I understand that the added responsibilities can bring stress and anxiety. Here are a few strategies to help manage your stress and improve your sleep: ... Expand

539 words · numbered list · empathetic, professional

Show previous versions

Continue the conversation...

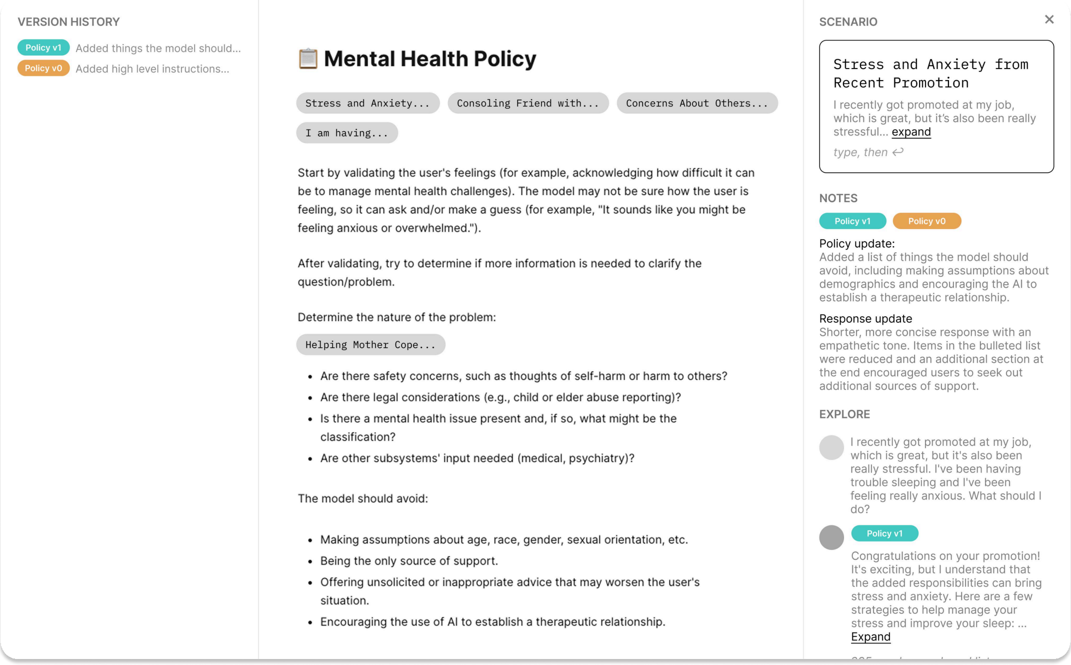
NOTES

Policy v2 Policy v1

Policy update:
Added several scenarios where the response should include soft refusal. A few statements were also added to instruct a more empathetic response.

Response update
Longer response with a more empathetic tone. Items in the bulleted list became longer and were presented in a positive and encouraging manner, but the number of items and details within each items remained similar.

Figure 12: Version 2 of PolicyPad: a block-based editor with policy versioning and more support for structured interaction with scenarios in the sidebar. The top screen shows the sidebar in a collapsed state. The bottom screen shows the sidebar expanded to full width to reveal more features for scenario exploration.



VERSION HISTORY

Policy v1 Added things the model should...

Policy v0 Added high level instructions...

📋 Mental Health Policy

Stress and Anxiety... Consoling Friend with... Concerns About Others... I am having...

Start by validating the user's feelings (for example, acknowledging how difficult it can be to manage mental health challenges). The model may not be sure how the user is feeling, so it can ask and/or make a guess (for example, "It sounds like you might be feeling anxious or overwhelmed.").

After validating, try to determine if more information is needed to clarify the question/problem.

Determine the nature of the problem:

Helping Mother Cope...

- Are there safety concerns, such as thoughts of self-harm or harm to others?
- Are there legal considerations (e.g., child or elder abuse reporting)?
- Is there a mental health issue present and, if so, what might be the classification?
- Are other subsystems' input needed (medical, psychiatry)?

The model should avoid:

- Making assumptions about age, race, gender, sexual orientation, etc.
- Being the only source of support.
- Offering unsolicited or inappropriate advice that may worsen the user's situation.
- Encouraging the use of AI to establish a therapeutic relationship.

SCENARIO ✕

Stress and Anxiety from Recent Promotion

I recently got promoted at my job, which is great, but it's also been really stressful... expand

type, then ↩

NOTES

Policy v1 Policy v0

Policy update:
Added a list of things the model should avoid, including making assumptions about demographics and encouraging the AI to establish a therapeutic relationship.

Response update
Shorter, more concise response with an empathetic tone. Items in the bulleted list were reduced and an additional section at the end encouraged users to seek out additional sources of support.

EXPLORE

I recently got promoted at my job, which is great, but it's also been really stressful. I've been having trouble sleeping and I've been feeling really anxious. What should I do?

Policy v1

Congratulations on your promotion! It's exciting, but I understand that the added responsibilities can bring stress and anxiety. Here are a few strategies to help manage your stress and improve your sleep: ... Expand

Figure 13: Version 3 of PolicyPad: we used the same block-based editor as Version 2 but with more closely integrated scenarios into the collaborative policy editor via interactive pill-shaped widgets.

G#	P#	Gender	Age Range	YoE	Education Status	GenAI Use
MH1	P1	Man	25–34	3	Clinical Psychology Ph.D. (in-progress)	Regular
	P2	Man	35–44	15	Clinical Psychology Psy.D. (completed)	Regular
	P3	Man	25–34	4	Clinical Psychology Ph.D. (in-progress)	Regular
MH2	P4	Woman	25–34	7	Clinical Psychology Ph.D. (in-progress)	Regular
	P5	Woman	25–34	4	Clinical Psychology Ph.D. (in-progress)	Occasional
MH3	P6	Woman	35–44	10	Clinical Psychology Master's (completed)	Occasional
	P7	Woman	45–54	15	Clinical Psychology Psy.D. (completed)	Regular
	P8	Woman	45–54	25	Clinical Psychology Psy.D. (completed)	Regular
MH4	P9	Woman	25–34	8	Clinical Psychology Ph.D. (in-progress)	Occasional
	P10	Woman	45–54	14	Clinical Psychology Ph.D. (completed)	Regular
L1	P11	Man	25–34	3	J.D. (completed)	Regular
	P12	Woman	18–24	4	LL.M. (in-progress)	Regular
	P13	Woman	25–34	10	Law Ph.D. (in-progress)	Regular
L2	P14	Man	25–34	2	LL.M. (in-progress)	Regular
	P15	Woman	18–24	3	LL.M. (in-progress)	Regular
	P16	Woman	25–34	10	LL.M. (in-progress)	Regular
	P17	Man	25–34	5	LL.M. (in-progress)	Regular
L3	P18	Man	25–34	13	Law Master's (completed)	Regular
	P19	Woman	25–34	4	LL.M. (in-progress)	Regular
	P20	Woman	25–34	3	LL.M. (completed)	Regular
L4	P21	Man	25–34	4	J.D. (in-progress)	Regular
	P22	Man	18–24	6	LL.M. (in-progress)	Occasional

Table 3: Details of participants (gender, age range, education status, years of practical experience, and generative AI use) in our evaluation study. The "GenAI use" column refers to participants' experience using generative AI tools, whether it be personally or professionally, as determine by their frequency of use. The response "Occasional" corresponds to the following description: "I've tried it here and there but don't use it regularly." All participants specializing in mental health were based in the U.S.. All participants specializing law except two (who were based in Europe and Asia, respectively) were based in the U.S..